THE UNIVERSITY *of* EDINBURGH

Deep Boltzmann Machines as Hierarchical Generative Models of Perceptual Inference in the Cortex

David P. Reichert

Doctor of Philosophy
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
2012

Abstract

The mammalian neocortex is integral to all aspects of cognition, in particular perception across all sensory modalities. Whether computational principles can be identified that would explain why the cortex is so versatile and capable of adapting to various inputs is not clear. One well-known hypothesis is that the cortex implements a generative model, actively synthesising internal explanations of the sensory input. This 'analysis by synthesis' could be instantiated in the top-down connections in the hierarchy of cortical regions, and allow the cortex to evaluate its internal model and thus learn good representations of sensory input over time. Few computational models however exist that implement these principles.

In this thesis, we investigate the deep Boltzmann machine (DBM) as a model of analysis by synthesis in the cortex, and demonstrate how three distinct perceptual phenomena can be interpreted in this light: visual hallucinations, bistable perception, and object-based attention. A common thread is that in all cases, the internally synthesised explanations go beyond, or deviate from, what is in the visual input. The DBM was recently introduced in machine learning, but combines several properties of interest for biological application. It constitutes a hierarchical generative model and carries both the semantics of a connectionist neural network and a probabilistic model. Thus, we can consider neuronal mechanisms but also (approximate) probabilistic inference, which has been proposed to underlie cortical processing, and contribute to the ongoing discussion concerning probabilistic or Bayesian models of cognition.

Concretely, making use of the model's capability to synthesise internal representations of sensory input, we model complex visual hallucinations resulting from loss of vision in Charles Bonnet syndrome. We demonstrate that homeostatic regulation of neuronal firing could be the underlying cause, reproduce various aspects of the syndrome, and examine a role for the neuromodulator acetylcholine. Next, we relate bistable perception to approximate, sampling-based probabilistic inference, and show how neuronal adaptation can be incorporated by providing a biological interpretation for a recently developed sampling algorithm. Finally, we explore how analysis by synthesis could be related to attentional feedback processing, employing the generative aspect of the DBM to implement a form of object-based attention.

We thus present a model that uniquely combines several computational principles (sampling, neural processing, unsupervised learning) and is general enough to uniquely address a range of distinct perceptual phenomena. The connection to machine learning ensures theoretical grounding and practical evaluation of the underlying principles. Our results lend further credence to the hypothesis of a generative model in the brain, and promise fruitful interaction between neuroscience and Deep Learning approaches.

Acknowledgements

It is done! Finally!

This process was not without its challenges. From the start, it was driven by intuitions and my vague ideas of how to 'solve the brain', and especially in the first year, my supervisors were not always quite convinced this was going somewhere... and understandably so, given the breadth of this project and its slightly unusual approach. Thankfully, the first results to materialise were enough to demonstrate the potential of this work.

First and foremost, I would thus like to thank my supervisors, Amos Storkey and Peggy Seriès. Amos and Peggy were always available for support, and gave me both the freedom to follow my ideas and the feedback needed to actually turn this into a coherent story. Thanks to Amos for giving me the chance to pursue this PhD, and to Peggy for filling the role of second supervisor, providing constant critical insight and input beyond what would have been required in her position. To acknowledge the support of my supervisors I will use the first person plural throughout this thesis.

I would like to thank my friends and fellow researchers at the Neuroinformatics DTC and iANC, in particular Nicolas Heess and Matthew Chalk, for helpful discussions. I thank John Tsotsos and Geoff Hinton for letting me spend several months in their labs in Toronto, and again Geoff for inspiring me with his work, without which this PhD would not have been possible in its current form.

Second-to-last and far from least, I thank my parents, who without exception always supported me. Thanks to my nieces and nephews for being so fond of their uncle despite his constant absence, and to my friends also outside the DTC who made my time in Edinburgh worthwhile; and to Karen for making the final write-up period of my PhD, despite all the stress involved, a blast.

Finally, I would like to acknowledge that this research was undertaken almost exclusively with free software developed by the open-source community, and benefited from code provided by other machine learning labs. Simulations were run on Linux, using Python and the SciPy scientific computing library (Jones et al., 2001), on which my model implementation drew extensively. The latter was initially based on deep belief net code provided by Hinton & Salakhutdinov (2006). To run my code on GPUs, I experimented with two Python modules, Theano (Bergstra et al., 2010) and Gnumpy (Tieleman, 2010), developed by machine learning labs in Montreal and Toronto, respectively.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*David P. Reichert*)

Contents

Chapter 1

Introduction

The mammalian neocortex plays a central role in virtually all aspects of cognition: perception across all sensory modalities, language and abstract thought, mental imagery, working and long-term memory, attention, motor control, planning and executive control, and so forth (for reviews, see e.g. Mesulam, 1998, Felleman & Van Essen, 1991). To a degree, these different functions map onto an organisation of cortex into distinct specialised areas, which can differ anatomically. At the same time, these areas all share cortical organisational principles, such as a layered architecture, and anatomical differences are likely an evolutionary adaption of the same original underlying neuronal circuitry to different kinds of inputs and outputs (Kaas, 2011). Indeed, at least within sensory cortex, the function of a cortical area is to a significant degree determined by the sensory inputs it receives during development. For example, what is normally auditory cortex can partially assume the role of a visual area if its inputs are rewired to be visual (Sur & Leamey, 2001). Thus, it seems warranted to ask whether one can identify general computational principles inherent to cortical processing and input representation, which allow the cortical machinery to flexibly deal with, or be adapted to deal with, many kinds of sensory data and many tasks involved in perception and beyond. What might these principles be, and how could they be captured by theoretical models?

One key property of cortical organisation is that cortical regions are arranged roughly hierarchically, or according to several parallel and interacting hierarchies. Such a hierarchy can be defined on anatomical grounds, because feedforward and feedback connectivity along it is asymmetrically realised in the layered cortical circuitry, and on physiological grounds, as neurons in higher areas appear to respond to more complex and abstract features of the sensory input than those in lower ones. In sensory cortex, this architecture is naturally interpreted as implementing a transformation from simple

to complex representation of sensory input across several processing stages. In ventral visual cortex for example, such a transformation occurs from simple features such as edges to surfaces and shapes to more abstract concepts such as object categories. At the same time, cortical processing is far from being simply a feedforward, sequential process from low level to high level or input to output. Feedback or top-down connectivity from higher to lower areas is vast, with most areas being reciprocally connected. These connections are implicated to mediate such processes as attention and top-down expectation, but their precise role remains poorly understood.

Hierarchical organisation is thus likely one key principle important for cortical processing. Less clear is how essential the feedback component is, whether it can in a first-order approximation be neglected, or whether it is an integral part of what constitutes cortical processing. Several theories that take the latter position are based on interrelated notions such as predictive coding, analysis by synthesis, adaptive resonance, generative models, and hierarchical Bayesian inference (e.g. Rao & Ballard, 1999; Yuille & Kersten, 2006; Carpenter & Grossberg, 1987; Lee & Mumford, 2003). What they have in common is a reverse transformation from high-level to low-level representation, mediated by feedback connections, to explain, predict, inform, generate, or reconstruct low-level representations of sensory data. Importantly, such a feedback communication can be used to *evaluate* both a current high-level interpretation during perception and, over time, the quality of high-level representations overall. If the cortex implements, in some sense, a model of the world on the basis of which sensory input can be predicted, then it can evaluate the quality of its model by analysing how well the predictions match the actual input. Hence, by improving its model to improve the predictions, such top-down processing could allow the cortex to *learn* a hierarchical representation in the first place, without supervision from an external teaching or reward signal. Powerful learning mechanisms, and in particular unsupervised learning mechanisms, could be further key ingredients for realising the versatility of the cortex and its capability for adaptation (e.g. Hinton, 2007).

Thus, an interesting hypothetical computational principle of cortical processing could be unsupervised learning in hierarchical representations. Unfortunately, so far there are few computational frameworks based on such principles, whether meant as biological models or not, and none that have been shown to deal with the challenging perceptual problems the cortex has to face, such as the many tasks associated with vision. As a starting point, it could be instructive to consider relevant developments in current artificial intelligence (AI) research, to see whether approaches there can serve as

models of cortical processing or at least as a point of comparison. Most relevant here is the branch of AI referred to as machine learning (ML), which emphasises learning from data over hard-coded algorithms, thus possibly similar in spirit to how cortex adapts itself to various forms of sensory data.

Two points of interaction between ML and biological modelling are classic connectionist neural networks and statistical/probabilistic models (another one is reinforcement learning, where external reward comes into play, e.g. Dayan & Daw, 2008). As the name implies, neural networks have long been considered as (admittedly overly simplistic) abstractions of neuronal processing in the brain. A highly relevant, recently emerging focus is 'Deep Learning' (Bengio, 2009), being concerned with neural network-like hierarchical (or 'deep') architectures that learn representations of data without supervision. This is realised by having each stage in the hierarchy learn to generate or reconstruct the activation patterns in the stage below. Hence, there seems potential to relate these models to hypothetical unsupervised hierarchical learning and feedback processing in the cortex.

Probabilistic approaches, on the other hand, have found significant interest in computational neuroscience and cognitive science under the heading of 'Bayesian models'. The term 'Bayesian' has various connotations, but for our purpose here it refers to a formal treatment of how to form beliefs about unknown variables characterising the state of the world from observed variables (sensory input), in the presence of uncertainty (due to a lack of knowledge, ambiguity, intrinsic noise of sensory signals, etc.). In particular, if these variables can be conceptualised as being hierarchically related to each other, then it might be possible to map inference in the resulting probabilistic model to hierarchical processing in the cortex (Lee & Mumford, 2003). However, so far it has proven difficult to translate idealised Bayesian models into the approximate algorithms the cortex would need to use, or to map probabilistic computations into neuronal circuits outside extremely idealised perceptual toy problems.

We argue that a model recently introduced in ML, the deep Boltzmann machine (DBM), is of particular relevance for reasoning about hierarchical processing in the cortex. The DBM is an instance of the Deep Learning approach, a neural network that learns hierarchical representations by utilising combined feedforward and feedback (i.e. recurrent) processing. Unlike related approaches, recurrent processing is employed not just during learning but also during inference ('perception'). It is a full generative model, meaning it can synthesise representations of input data even in the absence of the latter, allowing for a strong form of *analysis by synthesis* (e.g. Yuille & Kersten,

2006): the notion of perception as an active synthesis of an internal explanation of the sensory input that is evaluated against the latter, possibly via top-down connections in the cortex. And at the same time, the DBM also implements a *probabilistic* model of the data, and uses approximate, *sampling*-based inference methods that have indeed been proposed to underlie cortical representations and algorithms (e.g. Sanborn et al., 2010; Hoyer & Hyvärinen, 2003; Fiser et al., 2010). While Deep Learning approaches such as the DBM are often taken to be inspired by the brain (Bengio, 2009), the relevance of the DBM as a concrete model of processing in the brain has not been explored so far.

1.1 Overview over the thesis

In this thesis we studied the deep Boltzmann machine (DBM) as a model of generative processing and analysis by synthesis in the cortex. The aim was to show how several perceptual phenomena could be related to such processing, naturally emerging from the underlying computational principles in a model system that was not hand-designed for any one of the phenomena in particular, and to thus provide indirect evidence that these principles could play an important role in the cortex. Specifically, we modelled visual hallucinations, bistable perception, and object-based attention. Our results should motivate further fruitful interaction between computational neuroscience and ongoing machine learning research, in particular by taking into account the recently developed Deep Learning approaches. Due to the nature of the DBM as a probabilistic model, our results also contribute to the current trend in cognitive science and computational neuroscience to consider approximate probabilistic inference in the brain. Framed differently, two main goals of this work can be identified. The first is to model hierarchical generative processing in the cortex. The second is to address the issue of approximate probabilistic inference in the brain, which in particular relates to the first goal because it has been proposed that hierarchical cortical processing could be understood in terms of such inference (Lee & Mumford, 2003; Yuille & Kersten, 2006).

Each of the three perceptual phenomena modelled, to be summarised below, can be seen as emphasising a different key aspect of the model, though all aspects play a role in all cases. The work on hallucinations focuses on the synthesis of internal representations of sensory data. Bistable perception is explained in terms of approximate probabilistic inference, and this is thus the part where the issue of Bayesian approaches to cognition becomes most relevant. And finally, the work on object-based attention

examines in particular feedback processing in the cortical hierarchy, employing the generative component of the model for attentional selection.

1.1.1 Content

The range of subjects touched on in our work is very broad, from Deep Learning to probabilistic inference in the brain to attention. With space being limited, it is thus not possible to provide the extensive discussion each of these topics would in principle deserve. For the three perceptual phenomena modelled, we discuss them and selected literature in the respective result chapters. We make use of the remainder of this **introduction chapter** to establish a possible context for our work in terms of Bayesian or probabilistic models of cognition. This allows us the clarify some of the underlying concepts useful later as well as describe some issues surrounding these approaches, and also motivates further why our own approach is relevant.

In **Chapter 2**, the DBM and its technical aspects are introduced. The latter include the model formulation, its connection to related machine learning approaches, learning rules, and Markov chain Monte Carlo methods for inference. In **Chapter 3**, we then offer a novel perspective of the DBM as biological model, discussing issues such as its general biological relevance and plausibility, possible interpretations of the involved learning algorithms, and its conceptual status as a tentative model of the cortex. We also describe a procedure for decoding the internal states to allow modelling of perception. The relevance of the DBM as a biological model is then demonstrated more concretely in the subsequent result chapters, each of which is based on a published paper (Reichert et al., 2010, 2011a,b).

In **Chapter 4**, we relate the capability of the DBM to synthesise internal representations to visual hallucinations, specifically in Charles Bonnet syndrome, where complex hallucinations are caused by a loss of vision. We introduce biological homeostasis mechanisms in the model, and demonstrate that several qualitative aspects of the syndrome can be reproduced in simulation experiments. We also provide a novel model of the neuromodulator acetylcholine as influencing the balance between bottom-up and top-down flow of information in the cortical hierarchy.

The homeostatic mechanism underlying hallucinations can, on a faster timescale, also be interpreted as neuronal adaptation during ongoing perceptual inference. In **Chapter 5**, we show how such adaptation can be understood as enhancing probabilistic sampling. To this end, we give a biological interpretation to a recently introduced Boltz-

mann machine sampling algorithm known as rates-FPCD (Breuleux et al., 2011). This allows us to model the phenomenon of bistable perception, combining two seemingly alternative explanations for bistability based on either probabilistic sampling *or* adaptation. With the DBM as concrete 'neural' model we thus also bridge related high-level Bayesian models and low-level mechanistic ones. We model bistable perception from ambiguous images (the Necker cube) and binocular rivalry, and consider a potential involvement of spatial attention. This part of our work touches on several issues concerning (approximate) probabilistic inference in the brain, which we address in some detail in the discussion section of this chapter. In particular, we provide a detailed analysis of related probabilistic approaches to bistable perception in this light.

The final part of the results is concerned with attentional processing, presented in **Chapter 6**. Here, we relate recurrent inference in the DBM to theories of attention in the brain that posit that such processing serves to organise represented information according to what is and what is not relevant to the object in the focus of attention. In particular, in the hierarchical cortical architecture, higher stages might primarily represent the attended object only, and guide selection of relevant information in lower stages via feedback. We show how inference in the DBM relates to attentional processing, differentiate between object-based and spatial attention, and introduce novel sparsity-based suppressive mechanisms. We argue that making the connection between generative models and attentional processing can provide novel perspectives on attention in the brain as well as inspire new machine learning approaches that could cope with complex perceptual tasks in an effective and biologically plausible fashion. This is early work, and we discuss some of our attempts to extend it further.

Finally, in **Chapter 7**, we address any remaining issues. We discuss other generative models of cortical processing, suggest future work, and conclude with some overarching remarks, including some final comments on the issues surrounding probabilistic approaches to cognition.

1.2 Bayesian models of cognition: from ideal observers to cortical inference

Bayesian approaches have found prominence in the last two decades in cognitive and computational neuroscience (for reviews and introductions, see Knill & Pouget, 2004; Griffiths et al., 2008; Vilares & Kording, 2011). The Bayesian formalism allows for a

principled treatment of how beliefs about uncertain variables and events in the world should be formed on the basis of observed evidence in an act of inference. It has been suggested that Bayesian methods could be key to understanding processing in the brain, which itself might be Bayesian in some sense. In particular, as has been elaborated on at the beginning of this chapter, hierarchical processing in the cortex might correspond to Bayesian inference in an internal probabilistic model (Lee & Mumford, 2003). At the same time, these approaches have also been subject to critique, and there is currently a debate in the field concerned with their merit and conceptual foundations (Jones & Love, 2011a,b; Bowers & Davis, 2012a; Griffiths et al., 2012; Bowers & Davis, 2012b).[1] Key issues are claims to optimality (and whether they are made in the first place), and whether Bayesian models are meant as descriptions of the world, implying some form of ideal observer, or of psychological processes in the brain. Indeed, what makes a model 'Bayesian' might not be clear as such. In the broadest sense, Bayesian approaches could be simply characterised as those that use probability theory to model cognition.

Our own work connects at several points to other probabilistic or Bayesian approaches to cognition. While probabilistic inference is not the only perspective from which our work can be viewed, it does play a significant role, especially in the chapter on bistable perception (Chapter 5) and in the machine learning view on the DBM model. We thus make use of the remainder of this chapter to briefly introduce the underlying concepts and surrounding issues, as well as to provide a short characterisation of the spectrum of models that have been framed as 'Bayesian'. The purpose of our description is threefold: to explain the key concepts that will be relevant for discussing our model and results, to introduce the mindset behind other approaches that will be considered as related work, and lastly, to motivate further why our work is a timely contribution in light of the current debate.

1.2.1 The Bayesian formalism

At the heart of the Bayesian approach is to interpret data by evaluating different hypotheses that could explain it, and to use probability calculus to explicitly express and account for any uncertainty over the variables involved. The approach is often characterised by the application of *Bayes' rule*. Consider some random variables (or sets of variables) h and x, referring to events in or properties of the world. The probability for

[1] The discussion of Bowers & Davis, 2012a; Griffiths et al., 2012; Bowers & Davis, 2012b was published just as this thesis was being written. Many of the points touched on in this section are addressed there more extensively.

these variables to take on any particular combination of values is given by a *joint* distribution $P(x,h)$.[2] However, when only x is observed, then h needs to be inferred from that evidence. In the Bayesian context, the *hidden* or *latent* variable h would be taken to describe a hypothesis that in some sense explains the evidence. Example scenarios would be inferring a disease from observed symptoms, or inferring the presence and identity of objects in a visual scene from the observed image on the retina. As will be elaborated on later, this 'hypothesis' or explanation might also take a more implicit form, such as an internal distributed representation of sensory input.

Given that the observed evidence could be compatible with multiple hypotheses to various degrees, one seeks to infer the *posterior* probability distribution $P(h|x)$, which yields the probability of any hypothesis h given x was observed. From the definition of conditional probability, $P(x,h) = P(x|h)P(h)$ and equivalently $P(x,h) = P(h|x)P(x)$, one directly obtains Bayes' rule as

$$P(h|x) = \frac{P(x|h)P(h)}{P(x)}. \tag{1.1}$$

Here, the term $P(h)$ is the *prior* probability of a hypothesis h, i.e. the probability that is assigned to h before x is observed. The term $P(x|h)$ is the probability of the data under a given hypothesis, and thus a measure for how well a hypothesis is compatible with the data. Note that in computing the posterior, the data variable x is given and thus fixed, and the posterior is evaluated for different hypotheses h. Thus, in this context $P(x|h)$ is evaluated as a function of h with fixed x. This function is called the *likelihood* function. Finally, the term in the denominator serves as normalisation. According to the law of total probability, it can be computed as $P(x) = \sum_{h'} P(x|h')P(h')$, i.e. as a sum over the terms in the numerator for different h', meaning that all other hypotheses that could explain the data are taken into account when computing the posterior probability of a particular one.

Bayes' rule directly follows from the general mathematical properties of probability. In statistics, what makes an approach Bayesian is the way probabilities are *interpreted*, namely as means of formally quantifying degrees of beliefs that a proposition is true (e.g., what is the probability that it rains next Tuesday? What is the probability that this is a cat and not a tiger?). In contrast, according to the *frequentist* view, probabilities are only defined as (limit) frequencies of outcomes of well-defined random

[2] For brevity, we in the following generally gloss over the distinction between a random variable and its specific outcome and simply write x for either. More precise would be to differentiate between the variable x and any specific outcome $\bar{x}$, writing $P(x = \bar{x})$ for the probability for x to take on the value $\bar{x}$. Similarly, we take x to be discrete. The treatment of continuous valued variables is for the purpose here analogous (replacing distributions with probability *densities*, sums by integrals, etc.).

experiments (e.g. tossing a coin). These different views come with different classes of methods in statistics and an accompanying philosophical discussion. Within Bayesian approaches in statistics, one can further distinguish between objectivist and subjectivist views. Roughly speaking, the former contend that there are 'objective' ways of deriving Bayesian beliefs and priors (e.g. on the basis of consistency or rationality arguments; priors should be invariant under reparametrisation, have maximum uncertainty *a priori*, etc.). Usually the goal is to formulate priors that avoid arbitrary assumptions and have as little influence on the inference as possible, while still allowing for using the tools of the Bayesian formalism (Jordan, 2009). Subjectivists on the other hand take the position that assumptions made in defining a Bayesian model are inherently subjective; for an observer making such assumptions, Bayesian calculus is the optimal means of utilising them, but they do not need to be objectively derived in some sense. We here will not address further these philosophical aspects as they pertain to the field of statistics, though part of the conceptual issues might relate to the issues surrounding Bayesian approaches to cognition.

1.2.2 The structure and interpretation of the hypothesis space

In Bayesian models of cognition, Bayesian inference is meant to represent or relate to perceptual inference, where perception results from the brain interpreting sensory data in light of prior knowledge. The Bayesian approach is often described as formalising the notion of perception as unconscious inference of Helmholtz (e.g. as in Kersten et al., 2004).[3] There are general aspects of Bayesian calculus as such that can play a role in modelling cognition, such as the means by which priors and data are combined and general effects like explaining away of evidence (Chater et al., 2011). In any one modelling application, it is however usually the specific assumptions made about the structure of the hypothesis space, the nature of the underlying variables and quantitative relationships between them, that gives a Bayesian model its content. One way of characterising the hypothesis space, which has become standard in statistics, is using graphical models, which describe the dependency relations between the variables (e.g. Jordan, 2004). See Vilares & Kording (2011) for an overview of Bayesian models of cognition in such terms. Generally speaking, the resulting models can be quite complex, with rich interactions between variables to be observed, inferred, or taken into account

[3] To furthermore go from perception to *decisions* in a Bayesian framework, one needs to deal with the concept of utility as well. This is the subject of Bayesian decision theory, not to be discussed further here.

as nuisance variables (i.e. variables that are part of the model but not ultimately of interest). Thus, a Bayesian model can entail much more than simply 'applying Bayes' rule' (cf. Jones & Love, 2011a).

Often, the relationship between the variables is taken to be a causal one, where the latent variables to be inferred are in some sense 'causes' of the observed ones. In particular, the relation might be described in terms of a generative process (for example, latent variables could describe properties of objects in a visual scene, which in turn generate the 2D image on the retina). In the case of images, Kersten et al. (2004) also further characterise a generative model as 'strong' if samples can be produced from it that consistently look like the data in question, though this might require further elaboration. In general however, relationships between variables in a probabilistic model need not be causal. Indeed, the DBM model employed in this thesis is an example of a Markov random field or undirected graphical model, where dependencies between variables are usually meant to reflect mutual constraints rather than causality (for instance, adjacent pixels in an image are more likely to belong to the same object).

1.2.3 Characterising Bayesian models

From reviewing the literature, it appears to us that there is not necessarily a clear common definition of what makes a model of cognition 'Bayesian' (and the term has yet another connotation in machine learning, for example). Such models might or might not involve a generative component, reflect statistics of the environment, focus on an explicit application of Bayes' rule, or be concerned with questions of optimality. At minimum, what they have in common is a formal treatment of uncertainty, and it is in this broad sense that we use the term in this thesis.[4] It appears that a potential disagreement or confusion about what Bayesian models are 'all about' (Bowers & Davis, 2012b; Jones & Love, 2011b) is in part what underlies the current debate.

To further conceptual clarity, we introduce here briefly three dimensions according to which different types of Bayesian models can be distinguished, referring back to this scheme throughout this thesis to characterise and contrast with each other probabilistic models of cognition, including ours. The first dimension (perhaps a key point of contention according to Jones & Love, 2011b) is to differentiate between: *external* Bayesian models, meant as a description of (aspects of) the external world and what can in principle be inferred about it, implying an 'ideal observer' against which real

[4] In that sense 'Bayesian' is a synonym for 'probabilistic' as long as the latter is understood to refer to the Bayesian interpretation of probability.

observers can be compared; and, *internal* models, where a probabilistic model describes psychological constructs or neuronal representations, and inference algorithms describe cognitive and/or neuronal processes in the brain.[5] The second dimension then further characterises models of the internal kind according to whether they form a description of *high-level* psychological constructs, or of *low-level* mechanisms close to the neuronal implementation in the brain.

Lastly, for a third dimension one might want to distinguish between *conceptual* and *instrumental* models. The former are by design meant to explicitly describe a given perceptual inference problem, with variables that have clearly defined, explicit meaning attributed to them, such as the category of an object or its position, or a stimulus property controlled by an experimenter. With the term instrumental (cf. Schwitzgebel, 2011; Chakravartty, 2011) on the other hand we refer to approaches that use the formalism of probabilistic models to make sense of sensory data, or to describe how the brain might do it, in a more general and flexible fashion. Rather than explicitly capturing any one perceptual inference problem, such models are tools for the purpose of achieving a (behaviourally relevant) goal, such as flexibly discovering structure in, and learning useful representations of, sensory data.[6] The model to be employed in this the thesis, the DBM, would be an example of an instrumental model.

1.2.4 Ideal observers, optimality, rationality

In the following, we briefly review representative examples spanning the spectrum of Bayesian models of cognition. We begin with several approaches that have in common that they involve considerations of optimality, again an issue at the centre of the current debate.

In *ideal-observer analysis* (e.g. Geisler, 2003; Kersten et al., 2004), a Bayesian model is intended to be an objectively correct description (in so far as is possible) of

[5] For internal models, to be precise one might want to further distinguish between the internal model the brain might form about the world, the *scientific* model used by the scientist to describe that model in the brain, and the probabilistic (graphical) model that is part of the overall scientific model. We hope the meaning of the term 'internal model' will be clear from the respective context. For external models, note that they can still involve the brain, e.g. by asking what an ideal observer could in principle infer about an external variable from neuronal activity. However, the view would still be that of an external observer that reasons about the external world.

[6] A conceptual model can be external (describing what can in principle be inferred about a property of the world from data) or internal (describing high-level psychological constructs or low-level neuronal representations explicitly referring to a conceptual variable). An instrumental model, in the context of models of cognition, is likely only sensible as internal model. Overall, this terminology is only a suggestion. A further comparison with concepts in the literature would be useful, e.g. with the often cited three levels of explanation of Marr (Marr et al., 2010).

the relationship between conceptual variables capturing aspects of the external world and available sensory data, subject to incomplete knowledge and inherent noise of the sensory processes. Together with well characterised performance goals, exact inference in the model then implies an ideal observer that can take all available information into account to make optimal decisions. As Geisler (2003) writes, "The purpose of deriving an ideal observer is to determine the optimal performance in a task, given the physical properties of the environment and stimuli". The ideal observer can serve as a benchmark for comparison with the real performance of an organism, or as a starting point for more realistic models of the latter. An example of an ideal-observer analyses would be to consider how well in principle two noisy photon sources can be discriminated, given the statistical counts of photons emitted, or given neuronal firing patterns in the retina or thalamus. As Geisler notes (op. cit.), there is no reason to expect that *real* observers match the performance of ideal ones in general. Nevertheless, some authors do appear to report that there are actually many cases in the literature where human performance *does* match that of the ideal observer (e.g. Knill & Pouget, 2004). Even in how far this claim is actually made is controversial (Bowers & Davis, 2012a; Griffiths et al., 2012; Bowers & Davis, 2012b).

An influential example of the line of work referred to concerns *cue combination*, as described by Ernst & Banks (2002), Ernst & Bülthoff (2004). Here, subjects need to estimate a property of the environment by combining different sensory cues, for example estimating the size of an object using both vision and touch. Their performance is captured well by a Bayesian[7] model where estimates stemming from individual cues are weighted by their individual uncertainty to arrive at an integrated estimate. This is deemed "statistically optimal", defined by the authors to mean that the combination rule yields the estimate with the lowest variance given the variances of the estimates from the individual cues. It should be noted that here, the model is only concerned with the final estimates as reported by the subjects. The environment, sensory cues, and various stages of neuronal processing that lead up to what the subjects ultimately report are treated only implicitly. The final estimate is optimal only to the degree that it corresponds to the best one could do at the final decision stage, *if* one assumes a two stage decision process that makes a final estimate given the two initial estimates from individual cues.

Similar considerations apply more generally e.g. for the review by Knill & Pouget

[7]To be precise, it uses a maximum-likelihood estimation that can be seen as a special case of a Bayesian model when priors happen to be non-informative (Ernst & Bülthoff, 2004).

(2004), which states that there are "myriad ways in which human observers behave as optimal Bayesian observers". The authors do note that people are usually *not* ideal observers when measured according to the uncertainty entailed in the physical stimulus alone. Rather, "the real test of the Bayesian coding hypothesis is in whether the neural computations that result in perceptual judgements or motor behavior take into account the uncertainty in the information available at each stage of processing." Implicit here are the assumptions that processing in the brain can be separated into distinct stages of computation, and that one can then somehow disentangle the imperfect computations performed at one stage and the means by which the outcomes of these computations are subsequently used by the next stage. The distinction between taking into account uncertainty and doing it optimally by some definition might also not be completely clear here.

We consider two more influential examples of Bayesian approaches that involve notions of optimality. The first is Weiss et al. (2002), which is concerned with *biases in motion perception*. Using a Bayesian formulation, it shows that several visual illusions in human observers can be explained in a unified manner as reflecting a single prior for slow motion speeds in the world. Normatively, this prior could be an adaptation of the visual system to the statistics of motion in the world, biasing sensory perception when inputs are unreliable. The authors hence term these illusions "optimal percepts" of an "ideal observer", although they do not attempt to substantiate the relation to statistics of the environment or the optimality of the resulting inference. Rather, assumptions about the prior's functional form are justified on the grounds of tractability and the match to behavioural data. In a related subsequent study, Stocker & Simoncelli (2006) set out to measure the form of the prior employed by people more directly from psychophysical data. Altogether, if one takes Weiss et al.'s results to mean that people use the true statistics of the environment in an objectively optimal fashion, then the computational model is an external one, and in so far as people match the ideal performance, it *also* corresponds to an internal model characterising perceptual inference. On the other hand, the work by Stocker & Simoncelli might suggest an alternative, purely internal interpretation of the Bayesian model as simply characterising human biases in a formal way.[8] This bias might be useful even if it is not necessarily 'optimal'.

Finally, an approach that has been influential in Bayesian cognitive science is the *rational programme* by Anderson (1991). Like ideal-observer analysis, it is concerned with "what the optimal behavioral functions are" (op. cit.), given a theoretical characterisation of the environment and of what is being optimised. As Anderson writes,

[8] It should be noted that Stocker & Simoncelli still refer to an "optimal observer model".

"The structure of such a theory is concerned with the outside world rather than what is inside the head". Unlike ideal-observer analysis as described by Geisler (2003), the theory is described as being concerned not with individual perceptual tasks but with the general environment an organism evolved in, and appears to pose that 'rational' accounts are generally well-suited to shed light on cognition, rather than being primarily a point of comparison. The concrete model of Anderson (1991) is one of human category learning, essentially employing a probabilistic clustering algorithm that is a precursor of current nonparametric models in machine learning (Sanborn et al., 2010). Whether such a model can indeed be justified as objective description of the external world, or whether it should rather be seen as internal model of psychological constructs, is being debated (Jones & Love, 2011a, and commentaries).[9]

To summarise the above, several Bayesian approaches involve claims to optimality, but whether such claims are valid and what exactly they entail is not always clear and undisputed (Bowers & Davis, 2012a). Notions of optimality can differ, in particular, one of them might be (implicitly or explicitly) that performance is 'Bayes-optimal' if decisions fully align with an observer's assumptions (as specified in a probabilistic model), whether these assumptions are objectively correct or not.[10] However, beyond these issues, there might also be a consensus emerging from the current debate that more work is needed that applies or relates Bayesian models to psychological and neuronal processes (Jones & Love, 2011a; Griffiths et al., 2012). This leads us to Bayesian models which are more clearly framed as *internal*, to be considered next.

1.2.5 Approximate probabilistic inference and its neural substrate

Probabilistic approaches are not restricted to characterising ideal observers and theoretically optimal performance. In terms of high-level accounts, one position compatible with the 'rational programme' of Anderson (1991) is to keep the notion of a Bayesian model implying an ideal or rational observer, but to see the actual inference algorithm used by the brain as an approximation to the latter, framing psychological processes as "rational approximations" in 'rational process models' (Sanborn et al., 2010). Given

[9] In philosophy, it seems to be a common position to understand the concept of 'categories' as relating to the mind, and not (necessarily) to the external world (Thomasson, 2012).

[10] This is reminiscent of the subjectivist interpretation of Bayesian probability in statistics. There are further notions of optimality not addressed here, e.g. describing a generative model that matches the data distribution well as "statistically optimal" (Berkes et al., 2011, who at the same time equate their usage of the term to the ones discussed earlier), or referring to finding the maximum a posteriori hypothesis as optimal 'defined in a Bayesian manner' (Rao, 1999).

a complex Bayesian model, such an approximation is often necessary simply because exact inference is intractable, in general or under the constraints of the biological hardware. Other high-level Bayesian models do not refer to notions of optimality at all (e.g. Sundareswara & Schrater, 2008). The results to be presented in this thesis also relate to approximate probabilistic inference and approaches framed as rational process models, most directly the part on bistable perception (Chapter 5). We thus will discuss some concrete examples there.

On a lower level, in models of computational neuroscience, another avenue of enquiry has been whether the mathematical entities and computations of Bayesian calculus can be mapped explicitly onto neuronal representations and processing. This is a question that is somewhat independent of whether perceptual inference on a high level can be characterised well in Bayesian terms (see Whiteley, 2008, for an extensive discussion and review of the relevant literature). In that sense, there can be both a high-level or low-level interpretation of a given probabilistic model.

Various schemes have been proposed by which neurons might represent and compute with probability distributions (see e.g. Vilares & Kording, 2011). Maybe most influential have been approaches where the firing of a population of neurons explicitly codes for the full distribution over some perceptual variable of interest. So far however, it has proven challenging to extend such approaches beyond simple scenarios where the variables of interest are very low-dimensional, such as single feature dimensions (e.g. orientation of a Gabor patch) or binary decision variables. This might mirror the difficulty of characterising human performance on the behavioural level as Bayesian ideal observers outside of simple perceptual tasks (Whiteley, 2008). It is not clear how these coding schemes can deal with, for instance, the complex high-dimensional problems inherent to vision, and the multiple stages of rich sensory processing that precede human behavioural decisions.

Another perspective that has become prominent recently is to examine neuronal computations in terms of *approximate* inference and representations. Such approaches might be more suited to explain how the brain can cope with complex perceptual challenges, and form the neuroscience counterpart to the aforementioned 'rational process models' in cognitive science, with promising potential for interactions between the two. One example model mapping approximate probabilistic inference to the brain is the predictive coding model of Friston (Friston, 2008). Here, cortical populations are suggested to represent distributions parametrically (in terms of sufficient statistics such as means and variances/precisions), approximating the full model posterior with a simpler

distribution. Alternatively, it has been proposed both in computational neuroscience and cognitive science that probabilistic inference in the brain is sampling-based, termed the *Sampling Hypothesis* (Hoyer & Hyvärinen, 2003; Fiser et al., 2010; Daw & Courville, 2008; Levy et al., 2009; Vul et al., 2009; Sanborn et al., 2010; Vilares & Kording, 2011). Then, neuronal activity might encode a set of samples, i.e. points in the potentially high-dimensional variable space, that together represent a distribution approximately. The model we employ, the DBM, can be seen as an instance of a sampling-based probabilistic approach. We will thus address the question of sampling in the brain at several points throughout this thesis.

Finally, it has also been proposed that hierarchical processing in the cortex could be interpreted as inference in hierarchically structured probabilistic models. Lee & Mumford (2003) analysed various neuroscientific findings in this light, and also suggested that inference could be sampling-based. Moreover, hierarchical probabilistic inference might also offer a way of framing the notion of the cortex implementing analysis by synthesis (Yuille & Kersten, 2006, see also beginning of this chapter). According to this notion, perception involves synthesising an internal explanation of the sensory input that is assessed in terms of how well it accounts for the latter, perhaps by utilising top-down connections in the cortex. Such top-down or feedback processing might correspond to the generative component in a hierarchical probabilistic model.

1.2.6 The role of instrumental Bayesian models

Whether high-level accounts of psychological processes or low-level descriptions of neuronal implementations, *internal* probabilistic models are primarily concerned with capturing what is in the brain rather than the external world. In particular, *instrumental* models naturally fall into the category of internal models. Such approaches often use a probabilistic theoretical framework to describe how the brain might utilise and learn internal representations of sensory data and the external world. If such an approach is formulated as a generative model, then perception consists of inferring latent variables that might be (loosely) described as generative 'causes' or 'explanations' of sensory input. But as instrumental models, the exact meaning attributed to these variables might be beside the point as long as they serve achieving behavioural goals.

The DBM is an example of a probabilistic approach that we would categorise as instrumental, as are: sparse coding (Olshausen & Field, 1996), predictive coding (Rao & Ballard, 1999), arguably the (more qualitative) account of hierarchical Bayesian infer-

ence in the cortex of Lee & Mumford (2003), and several other probabilistic generative models of cortical processing (Rao, 1999; Dean, 2006; Murray & Kreutz-Delgado, 2007; George & Hawkins, 2009; Dura-Bernal et al., 2011). Notably, most of these approaches (all but sparse and predictive coding) are described as utilising 'Bayesian' frameworks to model perceptual processing, though the authors generally avoid making claims w.r.t. people or brains being themselves 'Bayesian' in some sense (cf. e.g. Knill & Pouget, 2004; Friston, 2009). Hence, usage of the term 'Bayesian' is not restricted to models of the external and/or conceptual type described earlier, although it is the latter that the current debate appears to focus on. We leave the discussion of the semantics of the DBM as instrumental model for until after it has been introduced in detail (specifically, Section 3.3), and the discussion of the models above for the final chapter so that we can relate them to our model and results (Section 7.1).

In concluding our brief overview over Bayesian approaches to cognition, we would argue that instrumental models are particularly suited for our purpose as stated at the beginning of this chapter. Our aim is to examine hierarchical generative processing in the cortex, and Bayesian frameworks, whether qualitative accounts or concrete computational models, have been proposed to elucidate on such processing. Specifically however, one of our key motivations is the possibility that generative processing might explain how the cortex can flexibly implement various aspects of perception and beyond, and perhaps how it flexibly learns to represent various sensory inputs in the first place. To work towards understanding this aspect of cortex, we thus need in particular not only models that are formulated as explicit description of one perceptual task, e.g. describing how a specific property of the environment can be inferred from sensory data and/or be explicitly encoded with neurons; rather, we need models that consider general flexible computational mechanisms with which structure in sensory data can be discovered and utilised. Arguably, this role is naturally filled by models that we have characterised as instrumental. Framing instrumental models in the same theoretical language as other approaches to cognition, namely that of probabilistic inference, can allow for a fruitful interaction.

The recent work of Tenenbaum et al. (2011) might indicate that a similar perspective is emerging in some high-level Bayesian approaches that might have been termed 'rational' in the past. Their work involves a hierarchical Bayesian model that can discover different forms of representational structures in vastly different kinds of data (e.g. features of animals, colours, locations of cities, supreme court decisions). Interpreting this model as objectively correct descriptions of the world and its generative processes

seems difficult, and it carries little *a priori* meaning. Consequently, contrasting their approach to earlier ones, the authors state that their model here targets a view of

> cognition as approximately optimal inference in probabilistic models defined over a learner's subjective and dynamically growing mental representations of the world's structure, rather than some objective and fixed world statistics.[11]

We would see this as an example of a shift away from interpreting high-level probabilistic models as objective descriptions of a perceptual task and its optimal solution, and towards interpreting them as internal models with instrumental semantics.

1.3 Conclusion

Our aim is to explore generative processing and analysis by synthesis as computational mechanisms in the cortex, as they could underlie its capability to adapt to a variety of sensory inputs and behavioural functions. Hierarchical Bayesian inference has been put forward as a theoretical framework in which such mechanisms could be understood (Lee & Mumford, 2003; Yuille & Kersten, 2006), and the Bayesian formalism could offer a unifying language for computational models in cognitive science and neuroscience. For our purpose, we are interested in particular in models that are not formulated as explicit probabilistic description of any one perceptual problem, where variables to be inferred have by design meaning assigned to them in terms of the external world, but rather flexible systems that can form general representations of the world by discovering structure in sensory data.

Designing such models has proven difficult even in machine learning, where there is no need for establishing a biological interpretation, and efforts are ongoing. Deep Learning approaches in particular are motivated by similar goals and are taken to be inspired by the brain (Bengio, 2009). In this thesis, we show that one Deep Learning approach, the DBM, might be particularly suited to be taken as model of cortical processing, being based on principles such as unsupervised learning, generative processing, and analysis by synthesis, and carrying both the semantics of a neural network and of a probabilistic model. It even relates to notions of approximate inference currently considered in computational approaches to cognition. In what sense the DBM is to be interpreted as a biological model needs clarification, and this issue will be addressed in Chapter 3, after the model has been introduced in detail in Chapter 2.

[11] This is an example where 'optimal' appears to be meant in a more 'subjectivist' sense.

Chapter 2

Deep Boltzmann machines

In search for the computational principles that could underlie cortical processing (Chapter 1), deep Boltzmann machines (DBMs) have a variety of properties of interest. In this chapter, after giving a brief overview below, we introduce the basic technical aspects of Boltzmann machines (BMs) and the DBM in Section 2.1, including interpretations as stochastic neural networks and probabilistic graphical models. Then we explain the various learning algorithms involved (Section 2.2). The status of the DBM as biological model will be addressed in Chapter 3.

DBMs are probabilistic, generative neural networks that learn to represent and generate data in an unsupervised fashion. They consist of several layers of neuronal units arranged in a hierarchy. The units fire stochastically, inducing a probability distribution over the network state, parametrised by the weight parameters, i.e. the connection strengths between units. Letting the neurons fire stochastically implements a (Markov chain Monte Carlo) sampling algorithm on that distribution. In light of our discussion of different ways a probabilistic or Bayesian model can be interpreted (Section 1.2.2), DBMs as potential models of representations and processing in the brain would be classified as *internal* and *instrumental*: the model is not derived as an objective description of any specific generative process or statistical relationship between variables in the external world, but rather formulated to capture how the brain might learn and use flexible probabilistic internal representations to explain a variety of data.

As probabilistic models, DBMs thus potentially relate to other Bayesian approaches as discussed in Section 1.2. At the same time, they share aspects with other connectionist-style neural networks that have found application in computational neuroscience and cognitive science. BMs are essentially stochastic versions of the memory networks called Hopfield nets (Hopfield, 1982). Unlike the latter, BMs also have hidden units

and thus not only memorise data patterns, but rather learn representations of that data. The hidden layers in a DBM compute distributed representations in several non-linear processing stages, as is the case in a classic feedforward neural network. However, whereas the latter is trained by providing desired output values (i.e., in a supervised fashion) and tuning the weights with the backpropagation algorithm (Rumelhart et al., 1986), DBMs can learn without supervision by approximately maximising the likelihood of the data under the generative model. Thus, what drives the learning is that the internal model is attempting to find representations of the data that enable it to generate the latter.

DBMs were introduced recently in machine learning (Salakhutdinov & Hinton, 2009), and have not been examined from a computational neuroscience point of view before.[1] DBMs are just a special case of BMs by virtue of their 'deep' architecture. BMs themselves were developed in the nineteen eighties (Ackley et al., 1985). The reason that DBMs have enjoyed recent interest in machine learning is not that the basic underlying model formulation of a BM has changed; rather, recent developments in learning algorithms have made it possible to effectively train these models, taking advantage of their deep structure to overcome earlier problems that made the application of BMs difficult. One of the key aspects of these new algorithms is that the layers in the hierarchy are initially trained successively one layer at a time, with each layer learning to generate the activations of the units in the layer below. Together with several related neural network type models, such as convolutional neural networks and deep versions of auto-encoders, DBMs are studied in the 'Deep Learning' branch of machine learning. Deep Learning has emerged in recent years in the wake of the development of these new learning techniques (Bengio, 2009), with application to a variety of problems from image processing to speech recognition (e.g. Hinton & Salakhutdinov, 2006; Nair & Hinton, 2009; Salakhutdinov & Hinton, 2009; Mohamed et al., 2012; Vincent et al., 2008).

2.1 Model formulation

A standard Boltzmann machine (BM) consists of binary neuronal units connected with symmetrical weights. The overall state of the network is described with a state vector $\mathbf{x}$, with unit i being on if $x_i = 1$ and off if $x_i = 0$, and the weights together form the weight

[1] Lee et al. (2008) compared learnt receptive fields in the closely related deep belief nets to receptive fields in visual cortex .

matrix $\mathbf{W}$ (with $w_{ij} = w_{ji}$, $w_{ii} = 0$, for all i,j). Each unit also has a bias parameter b_i, with can be seen as controlling the excitability of the unit. The probability for unit i to switch on is determined by the input z_i it receives from the other units it is connected to via the weights, with the bias functioning as a baseline input,

$$z_i = b_i + \sum_j w_{ij} x_j. \tag{2.1}$$

This probability to be on is then computed using a sigmoid (logistic) activation function:

$$P(x_i|\mathbf{x}_{\setminus i}) = \frac{1}{1+e^{-z_i}}, \tag{2.2}$$

where $\mathbf{x}_{\setminus i}$ denotes all unit states other than x_i. $P(x_i|\mathbf{x}_{\setminus i})$ is also called the activation (probability) of unit i.

When the BM is run for a long enough time, the probability of finding it in a given state will converge towards an equilibrium distribution. This distribution depends on what is called the *energy* of the state,[2]

$$E(\mathbf{x}) = -\sum_i b_i x_i - \sum_{i<j} x_i w_{ij} x_j. \tag{2.3}$$

The equilibrium distribution is then given as

$$P(\mathbf{x}) = \frac{1}{Z} e^{-E(\mathbf{x})} \tag{2.4}$$

$$= \frac{1}{Z} e^{\sum_i b_i x_i + \sum_{i<j} x_i w_{ij} x_j}, \tag{2.5}$$

where Z, the so-called partition function, serves as normalisation constant, $Z = \sum_{\mathbf{x}'} e^{-E(\mathbf{x}')}$.

E is called the energy for a good reason. The mathematical formalism used to describe the BM originates from statistical thermodynamics in physics. The Boltzmann model itself is equivalent to an Ising model used to describe ferromagnetic materials, where the individual 'units' correspond to the magnetic moments of individual atoms. These interact and orient themselves in parallel or anti-parallel, determining the overall energy of the system as in Eq. 2.3. The equilibrium distribution then is given by the Boltzmann distribution, which yields the probability of a state of any (classical) thermodynamic system with energy E and temperature T as

$$P(\mathbf{x}) = \frac{1}{Z(T)} e^{-E(\mathbf{x})/(kT)}, \tag{2.6}$$

(where k is the Boltzmann constant). Thus, states with lower energy are are more likely.

[2] Or in matrix form, $E(\mathbf{x}) = -\mathbf{b}^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{W}\mathbf{x}$, where the $\frac{1}{2}$ comes from counting each connection twice in the symmetric matrix $\mathbf{W}$.

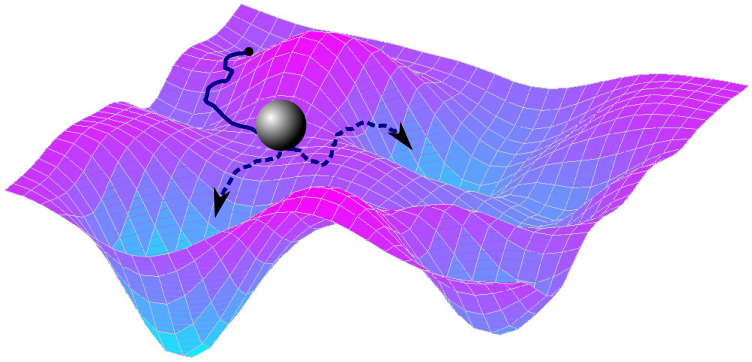


Figure 2.1: When a BM is run (corresponding to Gibbs sampling), the state of the system stochastically explores the model's energy landscape. The lower the energy of a state, the more likely it will be sampled.

The connection to statistical physics has lead to a variety of mathematical tools being imported from the latter to machine learning and statistics. For example, the temperature parameter T can be included in the probabilistic model and used to implement simulated annealing (Kirkpatrick et al., 1983), a method where the temperature is decreased gradually from a high initial value to search for the global energy minimum, which is in direct analogy to the physical process of annealing of materials. However, unlike with the Ising model in physics, the goal here is not to model a given physical system, but rather to use the BM to model data, matching the probability distribution defined by the model to that of the data by *learning* the weights and biases. This will involve making vectors that have high probability in the data have low energy in the BM. When the model is run, it stochastically explores the energy 'landscape', where it is more likely to transition to states with low energy (Figure 2.1).

Other types of neuronal units than binary are possible (see e.g. Welling et al., 2005). In this work, we also used softmax units to implement classification, which will be explained later in Section 3.5.

2.1.1 BMs as stochastic Hopfield nets with hidden units

A BM can be seen as a generalisation of the Hopfield network (Hopfield, 1982; MacKay, 2002), which has been used as a basic model of memory storage and recall in neuronal cell assemblies (Durstewitz et al., 2000). The Hopfield net is deterministic: its activation rule can be derived from the BM rule by taking the zero temperature limit (then a unit switches on or off deterministically depending on whether its input is greater or less

than zero). Used as a memory network, the task is to recall memorised patterns from corrupted or incomplete input patterns. Taking the latter as the initial state and letting the network then run freely, it will converge to a nearby state that is a local minimum in the energy. Thus, the learning rule for the Hopfield net needs to make the patterns to be memorised low energy stable states. The learning rule has a simple form that sets the weight between two units according to how correlated the activities of these units are over the N memory patterns, thus implementing a form of Hebbian learning ('neurons that fire together wire together'):[3]

$$w_{ij} \propto \sum_{n=1}^{N} x_i^{(n)} x_j^{(n)}. \tag{2.7}$$

A BM differs from a Hopfield network in two important ways. First, its activation rule is stochastic. Instead of finding the nearest minimum in the energy landscape, a BM performs a stochastic exploration of the latter. According to the Boltzmann distribution (Eq. 2.6), the state of the BM will be more likely to be found in a particular low energy state rather than a particular high energy one. However, in principle it can reach any point in the state space. The second difference is that unlike in a Hopfield net, some of the units in a BM can be treated as *hidden* units $\mathbf{h}$. Whereas the *visible* units $\mathbf{v}$ directly correspond to the data, the hidden units represent unobserved, latent variables. These hidden units enable the BM to learn aspects of the data that are not defined by pairwise correlations (cf. Eq.2.7). More generally, rather than just capturing correlations between visible units (e.g. pixels in an image) in the weights between them, hidden units can represent specific patterns or features in the visible units, and explicitly signal the presence or absence of the latter by virtue of their state. Thus, this opens up new possibilities for the BM over the Hopfield net; rather than just memorising patterns, it can learn internal representations of data. See Figure 2.2 on page 29 for an overview over different BM architectures, to be explained in detail below.

Both the probabilistic nature and the introduction of latent variables distinguish the BM from a Hopfield net and make it interesting from a probabilistic modelling point of view, as will be elaborated on in the following.

2.1.2 Modern machine learning view

Above, we conceptualised the BM as a neural network with a stochastic activation rule. When the network is run, the probability distribution over the state it is in approaches an

[3] This rule holds for binary states $x_i \in \{-1, 1\}$ rather than $x_i \in \{0, 1\}$.

equilibrium distribution, the network's Boltzmann distribution, in analogy to physical systems. This also allows for a more modern machine learning viewpoint of BMs as probabilistic graphical models of data. In this context, a BM is an instance of a Markov random field, which is a probabilistic graphical model whose independence relationships are captured by an undirected graph.[4] Rather than introducing BMs on the basis of the stochastic activation rule, one can instead start with the Boltzmann distribution (Eq. 2.4) as a definition of the model via the model's joint distribution over the random variables $\mathbf{x}$, and then derive what can be interpreted as an activation rule for each unit simply as conditional probability. By the definition of conditional probability, we have

$$P(x_i|\mathbf{x}_{\setminus i}) = \frac{P(x_i,\mathbf{x}_{\setminus i})}{P(\mathbf{x}_{\setminus i})} = \frac{P(x_i,\mathbf{x}_{\setminus i})}{\sum_{x_i'} P(x_i',\mathbf{x}_{\setminus i})}. \tag{2.8}$$

Due to the sum in the energy, the joint distribution factorises into a product. All terms not dependent on x_i can thus be pulled out of the sum in the denominator and cancelled with the respective factors in the numerator. It remains (biases omitted for clarity)

$$P(x_i|\mathbf{x}_{\setminus i}) = \frac{e^{\sum_j x_i w_{ij} x_j}}{\sum_{x_i'} e^{\sum_j x_i' w_{ij} x_j}} \qquad (\text{Note that } x_i' \in \{0,1\}) \tag{2.9}$$

$$= \frac{e^{\sum_j x_i w_{ij} x_j}}{e^0 + e^{\sum_j w_{ij} x_j}}. \tag{2.10}$$

For the activation rule, we specifically need the conditional probability for a unit to turn on, $x_i = 1$, which we plug in to arrive at

$$P(x_i = 1|\mathbf{x}_{\setminus i}) = \frac{e^{\sum_j w_{ij} x_j}}{1 + e^{\sum_j w_{ij} x_j}} \tag{2.11}$$

$$= \frac{1}{1 + e^{-\sum w_{ij} x_j}}, \tag{2.12}$$

which recovers Eq. 2.2. Thus, the BM is a special case of a probabilistic graphical model where the conditional probability of a variable has a 'neural' interpretation as an activation rule.

[4] See also Section 1.2.2. For a *directed* model, the probability distribution naturally factorises into conditional probabilities (e.g. $P(x,y,z) = (P(x|y)P(y|z)P(z))$, making them suitable to describe 'causal' dependencies. In an *undirected* model, the dependencies can be seen as soft constraints among groups of variables (e.g. Bishop, 2006) (e.g., two units tend to be on together). In that case, the distribution factorises according to (overlapping) groups ('cliques') of variables (e.g. $P(x,y,z) = (P(x,y)P(y,z))$. Note that for the BM, the joint (Eq. 2.5) factorises in this way due to the sum in the exponent. In general, some independence relationships can be captured both by undirected and directed models, others only by one of them (op. cit.).

2.1.3 Gibbs sampling and Markov chain Monte Carlo

Because running the BM leads (in the limit) to its state being distributed according to the Boltzmann distribution, doing so provides a means to produce *samples* from that distribution. Again, this procedure has a principled interpretation in the context of a more general theoretical framework employed in statistics and machine learning, namely that of Markov chain Monte Carlo (MCMC) methods (for an introduction, see e.g. Andrieu et al., 2003). Monte Carlo methods in general are applied to perform approximate computations with probability distributions (integration, optimisation, etc.) when exact ones are intractable.

Briefly, the idea behind MCMC methods is to produce samples from a target distribution P (here, the BM distribution) by designing a stochastic process that iteratively, over a sequence of 'time' steps, explores the state space in such a way that it spends more time where $P(\mathbf{x})$ is high. This involves defining a *transition operator* $T(\mathbf{x}', \mathbf{x})$ that determines the probability of transitioning from any current state $\mathbf{x}$ to any subsequent state $\mathbf{x}'$. The resulting process is an instance of a *Markov chain*, i.e. a sequence of random variables where the conditional probability of assuming any state in the next step only depends on the current state and not the states in the past (which is called the Markov property). The trick is then to design a transition operator such that a state obtained by running this Markov chain for a large enough number of time steps is indeed a sample from the target distribution P (at least in the limit of many time steps). It should be noted that *consecutive* samples produced this way are correlated, i.e. not independently drawn from the distribution.

Thus, no matter the initial state, after applying the transition many times we desire the resulting state to be distributed according to a distribution that converges to P as we take more steps. If the initial state is distributed according to a distribution π^0 (e.g. one that puts all mass on a single state), then after a single transition it will be distributed according to $T\pi^0$, after two according to $T^2\pi^{(0)}$, etc. We need this to converge to a unique fixed distribution, i.e. there needs to exist an *equilibrium distribution* for the chain, and that this equilibrium distribution equals P:

$$T^n\pi^0 \to P \quad \text{for} \quad n \to \infty, \text{ any } \pi^0. \tag{2.13}$$

In particular, this implies that P is a *stationary distribution* of the chain, i.e. that it is a fixed point of, or invariant under, T:

$$T \cdot P = P. \tag{2.14}$$

Intuitively, if we think of a population of 'particles' or 'walkers' in the space initially distributed according to π^0, then after applying T many times they will be approximately distributed according to P.

For an equilibrium distribution to exist at all, a Markov chain must fulfil certain properties (irreducibility, i.e. any state can be reached from any other, and aperiodicity, i.e. the chain cannot fall into fixed cycles). If that is established, then we need to ensure that P is invariant under T (Eq.2.14), implying that P is indeed the equilibrium distribution (due to the latter's uniqueness). A sufficient (though not necessary) condition for the invariance of P is to have a T that satisfies *detailed balance* w.r.t. P:

$$T(\mathbf{x},\mathbf{x}')P(\mathbf{x}') = T(\mathbf{x}',\mathbf{x})P(\mathbf{x}). \tag{2.15}$$

Again intuitively speaking, this ensures that P is not changing when the transition is applied by making sure that the 'flow' of particles from any state $\mathbf{x}$ to any state $\mathbf{x}'$ is exactly cancelled out by the reverse flow.

One particular MCMC method is Gibbs sampling. Here, the transition from one state to the next is performed by cycling through each of the individual variables (or groups thereof in blocked Gibbs sampling), sampling its state from its conditional probability (e.g. x_i from $P(x_i|\mathbf{x}_{\setminus i})$). The new overall state is obtained after all individual variables have been sampled in this way. It can be shown that the resulting Markov chain indeed fulfils detailed balance and has P as equilibrium distribution. Gibbs sampling thus is useful whenever it is difficult to sample from a target distribution directly, but comparatively easy to sample from its conditionals.

The crux is now that in the case of P being the BM distribution (Eq. 2.4), the process of updating the variables during Gibbs sampling according to the conditional probability in Eq. 2.2 is exactly equivalent to having the 'neurons' of the model 'fire' stochastically. Thus, machine learning provides a principled interpretation of the BM as a probabilistic graphical model and of its activation dynamics as implementing a MCMC algorithm. Because exact probabilistic calculations in the BM are often intractable,[5] in practice sampling is often necessary as a means of approximation, keeping with the picture of the BM as a stochastic neural network.

Finally, in anticipation of our work to be presented in Chapter 5, it should be noted that MCMC methods can themselves face problems. One is that sampling schemes

[5]Computations in BMs often involve sums over all the states of a subset of the units (e.g. for marginalising, i.e. summing out, some variables). In particular, the partition function Z (the normalisation constant) is notoriously difficult to compute exactly (or approximately), as it involves summation over *all* possible states, $Z = \sum_{\mathbf{x}} e^{-E(\mathbf{x})}$. For example, for a fully connected BM with 1000 binary units, this would be $2^{1000} \approx 10^{301}$ terms.

such as the one used by Gibbs sampling result in random walks in the state space, which can be a rather inefficient way of exploring it. Another one is that the different regions of high probability can be separated by low-probability ones, which are difficult to traverse. For Gibbs sampling specifically, consider a system of two binary variables whose states have very high probability when both variables are equal but a very low one otherwise ($P(1,1) = P(0,0) \gg P(1,0) = P(0,1)$). Starting from a high probability state, the Gibbs updates can reach the other high probability state only via by switching one variable at a time. If $P(1|0)$ or $P(0|1)$ are very low, then this might happen only very rarely. All these problems are exacerbated in high dimensions, and can keep the chain from reaching its equilibrium distribution in a practical amount of time steps. Such chains are said to be bad or slow at 'mixing'. In Chapter 5, we will examine sampling in a BM from a biological point of view in the context of bistable perception. In particular, we address the problem of bad mixing by giving a biological interpretation to a modified version of Gibbs sampling in the model.

2.1.4 The structure of the latent hypothesis space in BMs

The difficulty of computations in BMs depends on the connectivity structure in the weights. In this section we introduce the different architectures that have been proposed to make BMs more effective, including the deep Boltzmann machine, which is the focus of this thesis.

One important type of architecture that makes strong simplifying assumptions is the *restricted Boltzmann machine* (RBM, Smolensky, 1986) (Figure 2.2b). In the RBM, visible units $\mathbf{v}$ and hidden units $\mathbf{h}$ have no connections among themselves, but instead only between the two types (forming a bipartite graph). If one describes the RBM as consisting of two layers, an input layer of visible units and a hidden layer, then there are only connections between the layers but none within (no 'lateral' connections). The joint distribution can then be written as[6]

$$P(\mathbf{v},\mathbf{h}) = \frac{1}{Z} e^{\mathbf{v}^T\mathbf{W}\mathbf{h}+\mathbf{b}^T\mathbf{v}+\mathbf{c}^T\mathbf{h}} \tag{2.16}$$

(with $\mathbf{b}$ and $\mathbf{c}$ being the biases of the visible and hidden units, respectively). The reason that this structure simplifies the computations involved is that conditioned on the visible

[6]To be clear, for the full BM, $\mathbf{W}$ is symmetrical (with $w_{ii} = 0$ for all i) covering the connections between any two units. For the RBM, due to the bipartite connectivity, we can use a smaller non-symmetric matrix, with each row containing the weights of one visible unit and each column the weights of one hidden unit.

units, the hidden units are conditionally independent,[7] and vice versa for the visible units conditioned on the hidden units. Thus, the conditional distributions factorise, for example for the hidden units:

$$P(\mathbf{h}|\mathbf{v}) = \prod_i P(h_i|\mathbf{v}). \tag{2.17}$$

This makes this posterior easy to compute exactly, and also means that when sampling from it, all hidden units can be updated in parallel rather than sequentially (as the state of one hidden unit does not depend on the others). When sampling from the whole RBM, visible and hidden layers can be sampled in this manner in alternation.

The simplified structure of the RBM comes at the price of diminished expressive power of the model with regards to what kind of distributions it can represent. In particular, the posterior over the hidden states given the visible units (Eq. 2.17) is necessarily unimodal, simply because each of the factors is unimodal as a distribution over a binary variable. A multimodal posterior on the other hand will be key in our model of perceptual bistability in Chapter 5. A type of BM with more expressive power but still restricted structure is the *deep Boltzmann machine* (DBM, Figure 2.2c), which extends the RBM by adding several subsequent hidden layers, with connections only between adjacent layers and no lateral connections. The higher hidden layers can reintroduce dependencies between hidden units in the lower layers. At the same time, the special structure of the DBM can be still be exploited during learning. This is because the DBM can be seen as a stack of RBMs. The key idea behind learning in DBMs is then to (pre-)train the model layer-wise, one RBM at a a time (Section 2.2).

Given the current debate regarding 'Bayesian' models of cognition and the role of concepts such as priors and Bayes' rule (Section 1.2), it might be useful to note that in a BM, the prior over hidden units $P(\mathbf{h})$ is only implicitly defined, i.e. it would have to be computed by marginalising the full joint, $P(\mathbf{h}) = \sum_{\mathbf{v}} P(\mathbf{v},\mathbf{h})$. In particular, a BM is an example of a model where performing probabilistic inference, or deriving the equations for it, is not naturally formulated using Bayes' rule (in its default form that contains explicit expressions for prior and likelihood). In a RBM specifically, the posterior over the hidden units given the data is simply given by Eq. 2.17. In a DBM, approximate methods are usually necessary such as sampling or mean-field inference (the latter is

[7] Random variables x and y are conditionally independent given z iff $P(x,y|z) = P(x|z)P(y|z)$, or equivalently, iff $P(x|y,z) = P(x|z)$. For the RBM, the latter condition, $P(h_i|\mathbf{h}_{\setminus i},\mathbf{v}) = P(h_i|\mathbf{v})$ for any i, can be quickly verified by considering the conditional for the general BM (Eq. 2.2), setting any weights between h_i and the remaining hidden units to zero; or, it can simply be read off the graph of the RBM, given the independence semantics of an undirected graphical model.

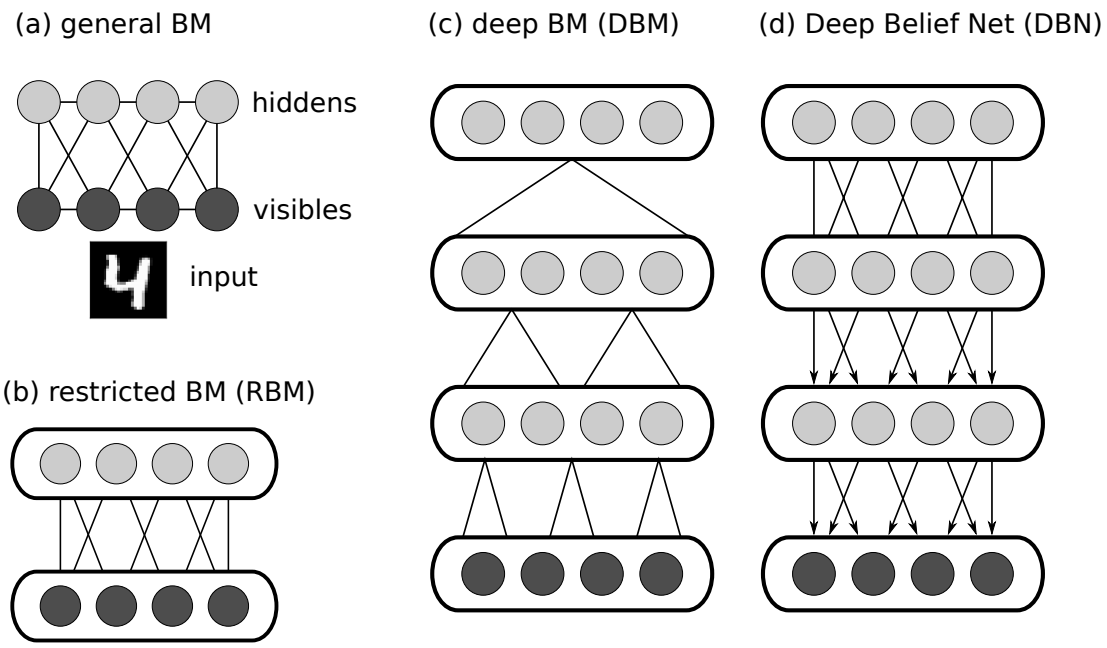


Figure 2.2: Boltzmann machines (BMs) with different architectures. A BM consists of a set of visible units which correspond to observed data (darker disks), such as pixels in an image, and a (possibly empty) set of hidden units. The latter represent latent variables that learn to capture higher-order correlations in the data. *(a)*: in a general BM, any pair of units can be connected (not all connections are drawn). *(b)*: in a restricted BM (RBM), visible and hidden units are separated into two layers, with no lateral connections within a layer. *(c)*: the deep BM (DBM) extends the RBM with additional hidden layers, again lacking lateral connections. As indicated by the depicted connections, in this work we often use localised receptive fields, which increase in size in higher layers (Section 3.6). (*d*): the deep belief net (DBN, Section 2.2.2), the first 'Deep Learning' model to be introduced, is a special case: it is *not* a BM. The topmost pair of layers forms a RBM, but connections below are directed.

explained in Section 2.2.2). In principle, Bayes' rule of course still holds, and could be applied if $P(\mathbf{h})$ were to be computed first, but in practice this is infeasible. In general, a simple and direct application of Bayes' rule will only occur where the involved terms (or at least the prior and likelihood, not necessarily the normalisation) are explicit in the model specification.

2.2 Learning

Given a set of N data points $\{\mathbf{v}^n\}_1^N$, we want to make a general BM consisting of visible and hidden units, $\mathbf{x} = (\mathbf{v}, \mathbf{h})$, a good model of the data by learning the parameters, i.e. weights and biases, $\Theta = \{\mathbf{W}, \mathbf{b}\}$. To this end, we employ maximum likelihood estimation of the parameters.

While the derivation of the BM learning procedures is somewhat involved, also due to a variety of approximations used, it should be noted that the resulting algorithms themselves are often surprisingly simple, and, as we we will argue, not implausible and possibly even interesting from a biological point of view.

Because only the visible units are observed, we want to match the likelihood $P(\mathbf{v})$ under the model to the data (for brevity we omit the dependency on Θ). The quantity to maximise is thus $\log P(\{\mathbf{v}^n\}_1^N)$ with the the hidden variables summed out, $\log P(\{\mathbf{v}^n\}_1^N) = \log \sum_{\mathbf{h}} P(\{\mathbf{v}^n\}_1^N, \mathbf{h})$, which is the cost function of the optimisation problem at hand (or rather, the negative cost function). To find the maximum, we compute the derivative w.r.t. to any parameter θ. Assuming the data is independent and identically distributed (i.i.d.), then

$$\frac{\partial}{\partial \theta} \log P(\{\mathbf{v}^n\}_1^N) = \frac{\partial}{\partial \theta} \log \prod_n P(\mathbf{v}^n) \tag{2.18}$$

$$= \sum_n \frac{\partial}{\partial \theta} \log P(\mathbf{v}^n), \tag{2.19}$$

hence we in the following do the calculations for an individual data point $\mathbf{v}^n$ and take the sum in the end. Before performing the differentiation, $\log P(\mathbf{v}^n)$ can be split further into two terms:

$$\log P(\mathbf{v}^n) = \log \sum_{\mathbf{h}} P(\mathbf{v}^n, \mathbf{h}) \tag{2.20}$$

$$= \log \sum_{\mathbf{h}} \frac{1}{Z} e^{-E(\mathbf{v}^n, \mathbf{h})} \tag{2.21}$$

$$= \log \sum_{\mathbf{h}} e^{-E(\mathbf{v}^n, \mathbf{h})} - \log Z. \tag{2.22}$$

In maximising $\log P(\mathbf{v}^n)$ by changing the parameters of the model, we seek to increase the first term, making the model assign higher unnormalised probability to the data point; and, we seek to decrease the second term, thus decreasing the total unnormalised probability assigned to *all* possible states ($Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}$). Together, the net effect is to increase the probability of the data under the model while decreasing the probability for other possible configurations, keeping the overall distribution normalised. Unfortunately, the intractability of the partition function Z will pose a challenge when computing its derivative.

As a side note, the negative of the first term in Eq. 2.22 is also called the free energy, $\mathcal{F}(\mathbf{v}) = -\log \sum_{\mathbf{h}} e^{-E(\mathbf{v}^n,\mathbf{h})}$. Note that with this definition, $P(\mathbf{v}) \propto e^{-\mathcal{F}(\mathbf{v})}$, in analogy to $P(\mathbf{v}^n,\mathbf{h}) \propto e^{-E(\mathbf{v}^n,\mathbf{h})}$. Hence, the free energy plays the role of the energy for the subsystem only defined over the visible variables. Learning in a BM with hidden units thus involves assigning low (in relative terms) free energy to the data.[8]

We now compute the derivatives of the terms in Eq. 2.22 w.r.t. parameter θ, starting with the first, data-dependent free energy term:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log\sum_{\mathbf{h}} e^{-E(\mathbf{v}^n,\mathbf{h})} &= \frac{1}{\sum_{\mathbf{h}} e^{-E(\mathbf{v}^n,\mathbf{h})}}\sum_{\mathbf{h}} e^{-E(\mathbf{v}^n,\mathbf{h})}\frac{\partial}{\partial\theta}(-E(\mathbf{v}^n,\mathbf{h})) && (2.23)\\
&= \frac{1}{Z\cdot P(\mathbf{v}^n)}\sum_{\mathbf{h}} Z\cdot P(\mathbf{v}^n,\mathbf{h})\frac{\partial}{\partial\theta}(-E(\mathbf{v}^n,\mathbf{h})) && (2.24)\\
&= \sum_{\mathbf{h}} \frac{P(\mathbf{v}^n,\mathbf{h})}{P(\mathbf{v}^n)}\frac{\partial}{\partial\theta}(-E(\mathbf{v}^n,\mathbf{h})) && (2.25)\\
&= \sum_{\mathbf{h}} P(\mathbf{h}|\mathbf{v}^n)\frac{\partial}{\partial\theta}(-E(\mathbf{v}^n,\mathbf{h})) && (2.26)\\
&= \left\langle \frac{\partial}{\partial\theta}(-E(\mathbf{v}^n,\mathbf{h}))\right\rangle_{P(\mathbf{h}|\mathbf{v}=\mathbf{v}^n)}. && (2.27)
\end{aligned}$$

As the notation in the last line indicates, the resulting term is the expected partial derivative of the negative energy under the model's posterior distribution, with the visible units clamped to data (the interpretation will become clearer once we plug in the concrete energy function of the BM later).

[8] Note that this free energy is *not* the same quantity as the *variational* free energy used in the context of variational Bayes, as is central to Friston's free energy framework (Friston & Stephan, 2007, also to be discussed in Section 2.2.2 and in the discussion chapter, Section 7.1.4).

The derivative of the second term, of the log partition function, is

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log Z &= \frac{1}{Z}\frac{\partial}{\partial\theta}\sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} && (2.28)\\
&= \frac{1}{Z}\sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}\frac{\partial}{\partial\theta}(-E(\mathbf{v},\mathbf{h})) && (2.29)\\
&= \sum_{\mathbf{v},\mathbf{h}} P(\mathbf{v},\mathbf{h})\frac{\partial}{\partial\theta}(-E(\mathbf{v},\mathbf{h}) && (2.30)\\
&= \left\langle \frac{\partial}{\partial\theta}(-E(\mathbf{v},\mathbf{h}))\right\rangle_{P(\mathbf{h},\mathbf{v})}. && (2.31)
\end{aligned}$$

Thus, we again obtain an expectation of the energy derivative, only this time it is taken over the whole model distribution. Combining both terms and taking the sum over all data points, we arrive at

$$\frac{\partial}{\partial\theta}\log P(\{\mathbf{v}^n\}_1^N) = \sum_n \left\{ \left\langle \frac{\partial}{\partial\theta}(-E(\mathbf{v}^n,\mathbf{h}))\right\rangle_{P(\mathbf{h}|\mathbf{v}=\mathbf{v}^n)} - \left\langle \frac{\partial}{\partial\theta}(-E(\mathbf{v},\mathbf{h}))\right\rangle_{P(\mathbf{h},\mathbf{v})} \right\}. \tag{2.32}$$

The maximum cannot be computed analytically in general. To find a (local) maximum, a gradient ascent algorithm is used, i.e. the parameters are changed iteratively, updating their current value by taking a step along the local gradient of the log likelihood so as to gradually increase it. Moreover, in practice one usually employs *stochastic gradient descent*: rather than taking the sum over all N data points (batch-learning), the gradient is approximated by computing it on only a single data point at a time, or a small subset of the data (a 'minibatch'), and then iterating over the data. One full iteration is also called an *epoch*. This restriction to few data points at a time makes dealing with even very large data sets feasible, and also allows for maybe more biologically relevant online-learning, where learning happens with continuously incoming sensory data rather than simultaneously over some large, fixed data set. In the following, the summation over N data points for each parameter update is thus to be understood as being over a small minibatch, or even a single data point.

The derivation of Eq. 2.32 holds true for any model that can be formulated in terms of an energy and a Boltzmann distribution, not just BMs.[9] For the BM in particular, we have $\frac{\partial}{\partial w_{ij}}(-E(\mathbf{x})) = x_i x_j$ for the weights and $\frac{\partial}{\partial b_i}(-E(\mathbf{x})) = x_i$ for the biases. Thus, as part of a gradient ascent we obtain the following update rules for the parameters of the

[9]Note that in principle, any probabilistic model can be written as $P(x) = \frac{1}{Z}f(x)$ for some function f (with $f(x) > 0$ on its support) and some normalisation constant $Z = \int \mathrm{d}x f(x)$, and thus can be formulated as $P(x) = \frac{1}{Z}e^{-E(x)}$ by defining the energy as $E(X) = -\log P(x)$. The argument here becomes relevant if this is a natural way of formulating the model and if Z is difficult to compute, as is often the case with undirected graphical models.

BM (with learning rate η):[10]

$$\theta_i \quad \hookleftarrow \quad \theta_i + \eta \frac{1}{N} \frac{\partial}{\partial \theta_i} \log P(\{\mathbf{v}^n\}_1^N), \tag{2.33}$$

where

$$\frac{\partial}{\partial w_{ij}} \log P(\{\mathbf{v}^n\}_1^N) \quad = \quad \sum_n \left\{ \left\langle x_i x_j \right\rangle_{P(\mathbf{h}|\mathbf{v}=\mathbf{v}^n)} - \left\langle x_i x_j \right\rangle_{P(\mathbf{h},\mathbf{v})} \right\} \tag{2.34}$$

$$\frac{\partial}{\partial b_i} \log P(\{\mathbf{v}^n\}_1^N) \quad = \quad \sum_n \left\{ \left\langle x_i \right\rangle_{P(\mathbf{h}|\mathbf{v}=\mathbf{v}^n)} - \left\langle x_i \right\rangle_{P(\mathbf{h},\mathbf{v})} \right\}. \tag{2.35}$$

Thus, at the core of the BM weight learning rule is computing correlations between any two connected units–a simple form of Hebbian learning in close analogy to learning in the Hopfield net (Eq. 2.7). Unfortunately, both expectations over the posterior and the full model distribution can be intractable in a general BM. In practice, these terms are thus approximated by sampling from the respective distributions, i.e. by replacing the expectation over the distribution by a sum over K samples drawn from it. For instance, $\left\langle x_i x_j \right\rangle_{P(\mathbf{h},\mathbf{v})} = \sum_{\mathbf{v},\mathbf{h}} P(\mathbf{h},\mathbf{v}) x_i x_j \approx \frac{1}{K} \sum_k x_i^k x_j^k$, where $x_i^k, x_j^k \sim P(\mathbf{h},\mathbf{v})$. These samples are generated by using Gibbs sampling, i.e. by running the BM as stochastic neural network.

Computing these two expectations by taking samples is also referred to as 'positive phase' and 'negative phase', respectively (Figure 2.3a). The positive phase can be seen as Hebbian learning while the BM receives input and infers hidden representations to interpret it, clamping the visible units to data points and sampling the hidden units. The negative phase on the other hand corresponds to Hebbian learning when samples are drawn from the whole model distribution (as implied by the current set of parameters), running the BM freely without input data, instead generating 'fantasy' data in the process. We will address the question of whether this algorithm could resemble waking and dreaming phases in the brain in Chapter 3.

Unfortunately, learning in general BMs turns out to be problematic. The reason is that mixing during sampling can be very slow. The BM needs to be sampled from for a long time to generate samples representative of the whole distribution, and this procedure needs to be repeated for *each* weight update in the gradient ascent. Moreover, the computed gradient can be noisy, being a difference between two stochastic estimates. Over the process of learning this can also lead to the parameters of the model effectively being pushed out of regions of high variance and effectively getting trapped where the variance is low (Hinton et al., 1998).

[10] The factor $\frac{1}{N}$ follows the convention used in the code of Hinton & Salakhutdinov (2006).

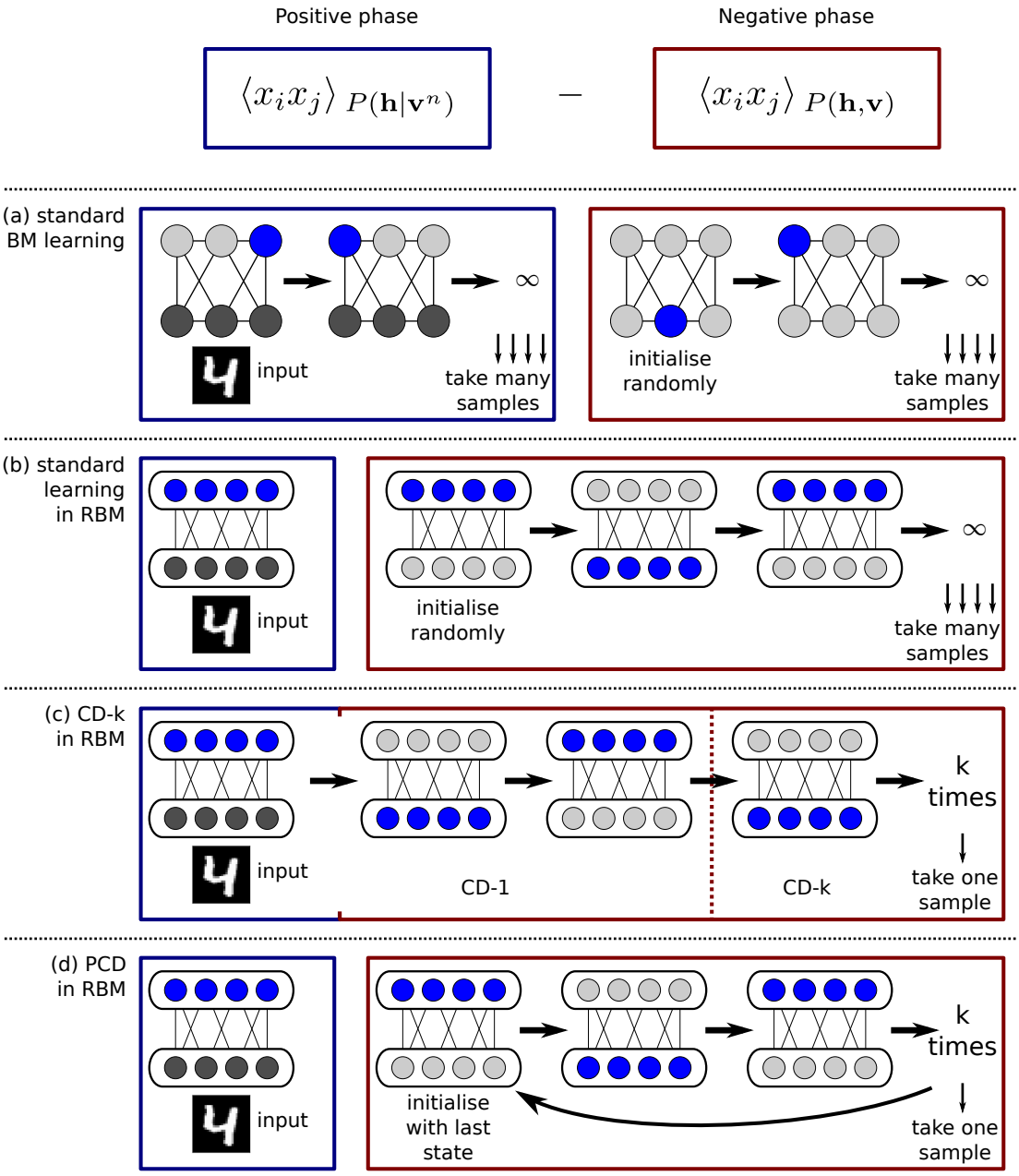


Figure 2.3: Caption see next page.

Figure 2.3: Learning in BMs is based on updating the weights according to the (approximate) gradient of the likelihood. Each weight update requires 'Hebbian', correlation-based learning between connected neurons in two phases: a positive phase where the visible units are set to the current data point $\mathbf{v}^n$ and the hidden units inferred, and a negative phase where all units are sampled freely. *(a)*: standard learning in fully connected BMs requires unfeasibly long sampling. *(b)*: in RBMs, the lack of lateral connections means that all units in each layer can be sampled in parallel. In the positive phase, the visible units are fixed and the posterior over the hidden units (the hidden activations) can be computed directly and exactly. Parallel sampling speeds up the negative phase, but convergence to equilibrium is still slow. *(c)*: In k-step contrastive divergence (CD-k), only a single sample of the RBM state is taken to compute the negative phase correlations. Starting from the inferred state of the positive phase, the visible units are sampled (obtaining a 'reconstructed' data point), followed by the hidden units. This is repeated k times, with the final sampled state taken for the negative correlations. *(d)*: Similarly, in persistent CD (PCD), few sampling steps are performed in the negative phase, and only the final sample contributes. However, the initial state is taken to be the final state from the negative phase in the previous weight update. This makes the Markov chain of the negative phase persistent across weight updates.

2.2.1 Learning in RBMs: (persistent) contrastive divergence

To render learning more effective, BMs with simpler connectivity structure like restricted Boltzmann machines (RBMs) can be used, which also allow for further approximate techniques exploiting this structure, such as the contrastive divergence algorithm. In a RBM, there are no lateral connections among visible or hidden units. The correlations in the weight update (Eq. 2.34) thus are always taken between a hidden and a visible unit, $\langle v_i h_j \rangle$. As the data-dependent waking term is an expectation over the posterior, and the latter is factorial in a RBM (Eq. 2.17), it can be readily computed:

$$\begin{aligned}
\sum_{\mathbf{h}} v_i h_j P(\mathbf{h}|\mathbf{v}^n) &= \sum_{\mathbf{h}} v_i^n h_j \prod_k P(h_k|\mathbf{v}^n) && (2.36)\\
&= \sum_{\mathbf{h}_{\setminus j}} \prod_{k\neq j} P(h_k|\mathbf{v}^n) \sum_{h_j} v_i^n h_j P(h_j|\mathbf{v}^n) && (2.37)\\
&= \sum_{h_j} v_i^n h_j P(h_j|\mathbf{v}^n) && (2.38)\\
&= v_i^n h_j P(h_j|\mathbf{v}^n)|_{h_j=1} + v_i^n h_j P(h_j|\mathbf{v}^n)|_{h_j=0} && (2.39)\\
&= v_i^n P(h_j = 1|\mathbf{v}^n). && (2.40)
\end{aligned}$$

Due to the conditional independence of the hidden units given the visible units, the positive phase for the update of weight w_{ij} thus only requires the posterior over h_j, with the states of the other hidden units rendered irrelevant. The expectation can be computed analytically simply by plugging in $P(h_j = 1|\mathbf{v}^n)$ (the activation of the unit) directly (Eq. 2.40), or by sampling from the posterior (Eq. 2.38) to keep with the interpretation of a stochastic neural network. Using the activations, we obtain for the weight updates (the treatment of the biases is analogous[11]):

$$\frac{\partial}{\partial w_{ij}} \log P(\mathbf{v}^n) = v_i^n P(h_j = 1|\mathbf{v}^n) - \langle v_i h_j \rangle_{P(\mathbf{h},\mathbf{v})}. \quad (2.41)$$

Unfortunately, the simplified RBM structure does not abolish the problems of the negative phase, where the expectation resulting from differentiating the log partition function is to be computed over the *full* model distribution. Sampling is sped up in RBMs, alternating between sampling hidden and visible layers in parallel, but overall convergence can still be slow (Figure 2.3b). To cope with this issue, an algorithm known as *contrastive divergence* was introduced (Hinton, 2002), which makes some further, rather strong approximations. For notational brevity, we will in the following assume that stochastic gradient ascent is performed one data point at a time ($N = 1$ in Eq. 2.33).

[11] The biases can also be absorbed into the weight matrix by introducing one additional unit that is set to be always on and which connects to all other units with weight vector $\mathbf{b}$.

For minibatches, the same computations just need to be performed for several data points, which can be done in parallel.

Contrastive divergence (CD) is motivated as using an approximation to the gradient for the weight update in RBMs, specifically by simplifying the problematic negative phase in the following way: rather than approximating the expectation over the model distribution by collecting a large amount of samples, CD uses only a *single* sample. Moreover, rather than running the Gibbs chain for a long time as would be necessary to generate a sample that is approximately from the equilibrium (i.e. model) distribution, the chain is truncated after a relatively small number of steps k (with the algorithm then called *k-step CD*, or *CD-k*). Most commonly even just one step is used, $k = 1$. Clearly, in such a scenario the initial state of the chain becomes highly relevant, contrary to the usual intent of MCMC, which is to 'forget' the initialisation and reach equilibrium instead. For CD, the chain is thus initialised to the *current* data point used in the positive phase, $\mathbf{v}^n$, and the hidden activation computed from the latter.

In summary, in CD the negative phase term is computed by taking the current data point and inferred hidden activations as starting point, and then sampling the RBM freely for several steps (alternating between sampling the visible layer and hidden layer as before). In particular, because the initialisation corresponds exactly to what is computed in the positive phase (Eq. 2.41), the positive and negative phases can be seen as two subsequent phases of one procedure (Figure 2.3c): in the positive phase, the posterior activations are inferred from the current input, and correlations computed between visible and hidden units contribute to the weight update. In the negative phase, the RBM is subsequently run freely for k further steps, starting by sampling the visible units from the hidden activations inferred in the positive phase. It is the correlations computed in the final step that then constitute the contributions of the negative phase.

CD, and especially CD-1, seem rather crude approximations to the true gradient (Eq. 2.41) (Hinton, 2010b). An alternative viewpoint suggests to interpret CD as being related to reconstruction error driven learning (Bengio & Delalleau, 2009). In CD-1, the positive phase consists of inferring the hidden representations, and the negative phase then computes a reconstructed data point from the latter, as well as another hidden code from the reconstruction. Note that if the reconstruction is perfect, then the resulting contributions to the weight update cancel out (on average). What CD-1 does *not* do is to ever explore the model distribution in regions that are further away from the data, as would be the case if the negative phase expectation were computed over the whole distribution. In particular, this means that when the model assigns probability mass

to regions of visible states not close to the data, these wrongly represented 'spurious modes' will not be found and eliminated by CD learning. Similarly, the negative phase contribution to the gradient can effectively lead to energy barriers being accumulated around the data points. Empirically (Desjardins et al., 2010; Breuleux et al., 2011, and our own experiments), CD is thus often found to be sufficient to learn features on data that allow for hidden codes useful for reconstructing input, but the resulting RBM is less capable as a generative model. For instance, it might need to be initialised with a data point to avoid falling into nonsensical spurious modes, and it can exhibit bad mixing, i.e. it needs to be run for a long time to traverse between modes of the distribution corresponding for example to different image categories the data was drawn from. Intuitively, it appears that these problems can be related to a lack of exploration of the model distribution away from the immediate vicinity of data during CD learning.

A more recent learning algorithm that improves on CD when it comes to exploration in the negative phase while keeping with its low computational complexity is *persistent contrastive divergence* (PCD, Tieleman, 2008; Younes, 1989). PCD has in common with CD that only a small number of samples is ultimately used in the negative phase (in practice, one per data point as with CD), and that the chain is run only for few iterations per update. Unlike in CD however, in PCD the negative phase is again decoupled from the positive phase. The chain is persistent in the sense that is continuously run across weight updates, rather than being reinitialised each time (Figure 2.3d). If the change in model parameters is slow compared with the mixing speed of the chain (e.g. when a very small learning rate is used), then the samples produced will be approximately from the current model distribution (albeit correlated if only few steps are taken between weight updates). In practice however, PCD works with much higher learning rates than expected. This has to do with dynamic effects on the RBM energy landscape (Tieleman & Hinton, 2009). We will return to this issue in the chapter on perceptual bistability (Chapter 5), showing how PCD and similar algorithms can be interpreted as being related to neuronal adaptation.

2.2.2 Learning deep architectures

The development of CD as effective training algorithm for RBMs made it possible to train more complex, 'deep' architectures, such as deep belief nets and deep Boltzmann machines (DBMs), which have multiple hidden layers lacking lateral connections. The key idea is to train the models one layer at a time, treating each pair of adjacent layers as

its own RBM. Other 'Deep Learning' architectures are non-generative models and for example based on auto-encoders (which are neural networks trained by reconstructing the input) rather than BMs, but similarly use layer-wise training. These approaches thus avoid the difficult optimisation problem posed by learning the full model directly. The layer-wise training can also be seen as *pre-training*, using it to initialise the full model in a good region of parameter space, then employing learning in the full model to further *fine-tune* it. Deep Learning approaches hence learn representations based on several stages of non-linear processing, and unlike traditional feedforward neural networks, do so in an unsupervised fashion. For this reason they have also been promoted in semi-supervised settings (Bengio, 2009), where the task does consist of predicting some labels (such as object classes in images), but only relatively few labels are available for training. The unsupervised learning makes it possible to utilise the unlabelled data to acquire knowledge about its inherent structure by learning useful features, then use the few labels available to further fine-tune the model in a supervised fashion.

Deep Learning was introduced with the deep belief net (DBN) (Hinton & Salakhutdinov, 2006), which we comment on briefly. As is the case with a DBM, a DBN is pre-trained by stacking RBMs. However, unlike a DBM the resulting full model constituting a DBN is *not* a BM. Rather, it consists of a RBM formed by the topmost and penultimate hidden layers, and a directed network below (Figure 2.2d on page 29). Generating samples from the DBN consists of first drawing a sample from the top RBM by Gibbs sampling, and then sampling each layer below in a single top-down pass. One reason for this somewhat peculiar architectural choice for the DBN is that this probabilistic model allows for a formal justification for why adding and learning consecutive hidden layers improves the model, at least for a particular choice of architecture where each subsequent RBM is initialised by transposing the weights of the layer below. However, in practice this specific choice is not actually commonly made (Bengio, 2009), nor does the argument hold for the DB*M*, the model used in our work. We thus do not further elaborate on it here.

Layer-wise pre-training of DBMs

The basic procedure for training a DBM layer-wise is simple (Salakhutdinov & Hinton, 2009). The first two layers consist of the visible units and the first hidden layer. Together they form a RBM, which is trained as usual on the data using for example CD or PCD. After the weights of this first RBM have been learned, a second hidden layer is added, with the pair of hidden layers forming the next RBM. This second RBM is

now trained using the activations (or samples) of the first hidden layer as data, where these activations are computed by inferring a set of hidden activations from the original training data. Subsequent hidden layers are added analogously.

In the resulting DBM, each of the intermediate hidden layers receives input from two adjacent layers. During training however, a newly added hidden layer is trained only with input from below. To compensate for this discrepancy, intermediate weights are simply halved when composing the DBM. A special case is the first RBM (and the final one, treated analogously). For the first hidden layer, bottom-up inputs should be halved as before. However, for the lowest layer, the visible layer, there is no need for halving the inputs, as they receive only top-down input in both pre-training and in the full DBM. The first set of weights can thus not simply be halved when composing the DBM. Instead, a trick can be used (Figure 2.4) where the first RBM is trained with duplicated visible units (always setting both sets to the same data points) and duplicated weights connecting both sets of visible units to the first hidden layer. The weights are tied, i.e. constrained to be the same, updating them together during learning. When composing the DBM, only one of the duplicated sets of visible units and weights is kept. This has the desired effect of halving the bottom-up input to the first hidden layer, while keeping the top-down input to that set of visible units unchanged.[12]

Fine-tuning of the full DBM

When composed, the DBM can be fine-tuned further by training it as a general BM, making use of the initialisation of the parameters obtained in pre-training. When they introduced DBMs, Salakhutdinov & Hinton (2009) suggested the following procedure for computing the weight updates (Eq. 2.34). For the negative phase expectation, PCD can be used, sampling from the model freely in a persistent chain (as with a RBM, units in any one layer can be sampled in parallel). For the positive phase, the posterior over the hidden units in the DBM no longer has the simple factorial form it has in the RBM, thus the expectation cannot be computed exactly anymore. Instead of sampling, Salakhutdinov & Hinton suggest using fast variational mean-field inference to approximate the posterior over the hidden states. As we will see, this method has yet again a

[12] In our experiments, we actually made a further simplification: instead of using a duplicated set of visible units, we simply used a multiplicative factor of 2 whenever computing bottom-up inputs to the first hidden layer during pre-training. During the positive phase, when the visible units are clamped to data, these two methods are equivalent. However, this is not the case in the negative phase, where duplicated visible units can be sampled independently. Nevertheless, this procedure worked for us in practice.

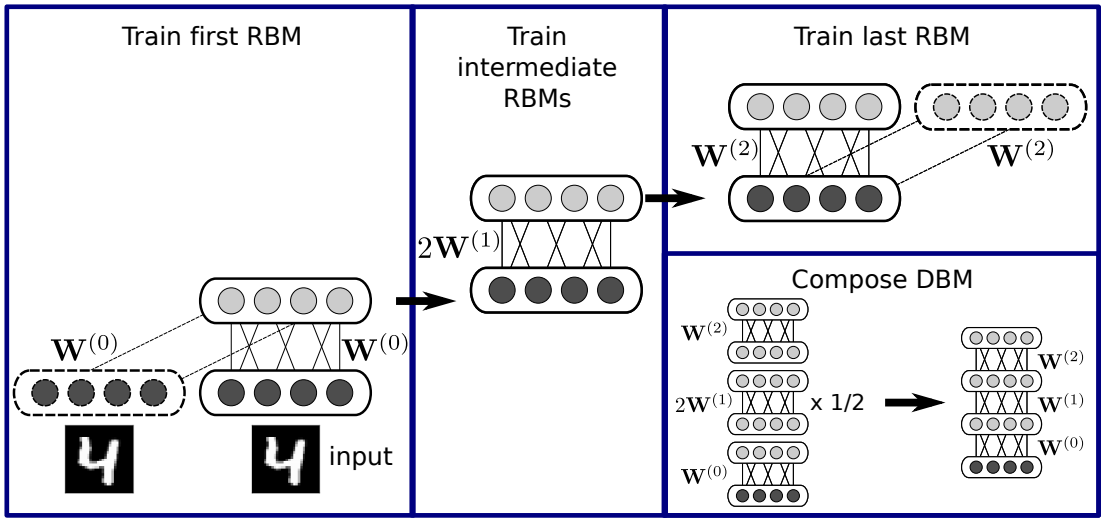


Figure 2.4: A sketch of DBM pre-training for three hidden layers. The basic idea is to train the DBM layer-wise, one RBM at a time, but there is a subtlety concerning how the final model is composed so that layers do not receive twice the amount of input in the end (see main text for explanation). Concretely, when training the first RBM and the last one, the visible layer or the ultimate hidden layer are duplicated, respectively, and weights to the duplicates are tied. When composing the full DBM (lower right), these duplicate weights are discarded; weights between intermediate hidden layers are simply halved.

simple, neural-network type interpretation. It should be noted that for our purposes, it turned out that we did not actually need to fine-tune the DBM. We describe the method here anyway for the sake of completeness, and because it also plays a role in other approaches to be discussed.

Briefly, in variational inference (e.g. MacKay, 2002; Wainwright & Jordan, 2007), given some data $\mathbf{v}$, the posterior $P(\mathbf{h}|\mathbf{v})$ over unobserved variables $\mathbf{h}$ is approximated with a simpler distribution $Q(\mathbf{h})$, termed the variational distribution. $Q(\mathbf{h})$ is matched to $P(\mathbf{h}|\mathbf{v})$ by optimising Q's parameters, minimising a quantity known as the variational free energy (which is closely related to, and implicitly minimises, the KL-divergence between the two distributions). Note that this optimisation is carried out for given, fixed observations $\mathbf{v}$. Computing the posterior for a different instances $\mathbf{v}'$ requires repeating the process of matching $Q(\mathbf{h})$ to $P(\mathbf{h}|\mathbf{v}')$.

In what is known as naive *mean-field* approach, the variational distribution is chosen to be a simple factorial distribution, $Q(\mathbf{h}) = \prod_i Q_i(h_i)$. In the DBM the variables are binary, thus each of the Q_i can be parametrised with a single parameter $\mu_i = Q_i(h_i = 1)$. As is often the case in variational inference, it turns out that the resulting optimisation problem can be solved iteratively, repeatedly updating the μ_i until convergence to a local optimum. The update rule for μ_i given the current values of the other parameters $\mu_{\setminus i}$ can be derived to consist of setting μ_i as the activation of h_i (Eq. 2.2 on page 21), to be computed from an input with the states of the other hidden units replaced with $\mu_{\setminus i}$, and then iterating through the other hidden units. Essentially, computing the mean-field posterior in a DBM corresponds to running it as a *deterministic* neural network, where, rather than sampling the units given their activations (i.e. $P(x_i = 1|\mathbf{x}_{\setminus i})$ for unit i), the activations themselves are propagated in the network until convergence.

The final state of the network defines the approximate posterior $Q(\mathbf{h})$. With the latter, the expectation term of the positive phase for weight w_{ij} then evaluates simply as $\mu_i\mu_j$ as inferred for each data point (or $v_i\mu_j$ for connections to a visible unit), i.e. as correlation between units computed using the mean-field activations. To summarise, fine-tuning a DBM can be done by using mean-field to compute the correlations of the positive phase, and running a persistent chain for the correlations in the negative one. Salakhutdinov & Hinton (2009) do not report whether simply sampling in the positive phase instead might work reasonably well, too, perhaps benefiting from the good pre-training initialisation of the parameters.

As a final note, we point out that the variational inference used in the context of DBM learning is a case of free energy minimisation as described in the framework of

Friston & Stephan (2007) and Friston (2009, 2010) (the 'free-energy principle'). Indeed, the concrete model of inference in the cortical hierarchy proposed there also employs a mean-field approximation to define the variational distribution Q. In that case the distribution P whose posterior is to be approximated is defined as a dynamic state space model in generalised coordinates of motion. In our case, it is given by the Boltzmann distribution for the DBM. Notably, Friston's general framework applies these concepts in contexts that reach far beyond the concrete hierarchical dynamical model, in ways that might deserve a further critical analysis. This is however beyond the scope of this thesis (we return to this subject very briefly in the final chapter, Section 7.1.4).

Chapter 3

DBMs as biological models

The main argument for the relevance of DBMs as models of biological, in particular cortical, processing will be delivered by reporting our results in Chapters 4-6. At this point, we can already make some general comments on the biological plausibility of the model and possible interpretations of its learning and inference mechanisms, and this is the subject of this chapter. In the next two sections, we focus on the overall properties of the DBM and its conceptual status as a biological model. Subsequently, in Section 3.3, we discuss the interpretation of the hidden variables and how it compares to other probabilistic models. In Section 3.4, we suggest relating the involved learning algorithms to predictive coding, dreams, and hierarchical development of the cortex. We then describe a method for decoding the internal state that enables us to use the DBM as model of perceptual phenomena, in Section 3.5. Finally, we briefly discuss the specifics of the model setup and training procedures used throughout this work in Section 3.6.

3.1 Biological plausibility and relevance

A first hurdle for approaching a DBM as computational neuroscience model might be the rather technical machine learning context it and related approaches, such as deep belief nets, were introduced in (Salakhutdinov & Hinton, 2009; Hinton & Salakhutdinov, 2006; Bengio, 2009). However, as we have seen in the last chapter, the aspects of the DBM that correspond to more general statistical methods all have a 'neural' interpretation in the case of the DBM. The model can be seen as a Markov random field with latent variables, where inference is performed by Gibbs sampling, or as a neural network with stochastically firing neurons. Learning involves maximising the likelihood

of the model parameters given data, which is implemented with gradient ascent, but the resulting update rules are simply a form of Hebbian (and anti-Hebbian) learning, measuring correlated activity between units. For training the full DBM, variational inference can be used, which in the form of a naive mean-field approximation results in replacing the stochastic activation rule for the neuronal units with a deterministic one.

Moreover, the connection to statistical methods is also an asset, as it allows for a comparison of the DBM to other machine learning approaches using principled theoretical tools. The interest of the machine learning community furthermore promises further developments in the future that could similarly have a relevance for computational neuroscience (e.g. Lee et al., 2008; Ranzato et al., 2010, to be commented on in the discussion chapter, Section 7.2).

As a model of neuronal processing in the cortex, a DBM is certainly an idealisation at a high level of abstraction. It lacks many essential and potentially functionally important biological details, from spikes to dendritic computations. As discussed above, none of its features seems inherently implausible to be implemented with biological mechanisms, but the correspondence to the real cortex is somewhat vague. For example, the hierarchical architecture of deep networks, involving several stages of nonlinear transformations, is often seen (Bengio, 2009; Hinton, 2010a) in analogy to the hierarchical organisation of the cortical regions. However, even within a cortical region there is of course a rich circuitry of cortical layers and lateral connections that is missing in the DBM, though the hierarchy of hidden layers might also play the role of lateral interactions. Similarly, the 'neurons' in a DBM are a simplified idealisation of true cortical neurons, sharing with them only core aspects, such as information integration through weighted channels ('synapses'), and realistic complex spiking behaviour is modelled only as simple on/off states. Finally, the symmetry of the connections between the neurons is also an idealisation. As in a Hopfield network, this makes it possible to implement the simplest form of Hebbian learning based on pairwise correlations, and to theoretically treat the attractor properties of the network using its energy function (MacKay, 2002). From a biological viewpoint, reciprocal connections between neuronal populations in the cortex are common, including between regions in the cortical hierarchy (Felleman & Van Essen, 1991). But of course, symmetric weights are a strong simplification, in a sense merging together feedforward and feedback connections, which are clearly differentiated anatomically (and possibly functionally) in the cortex.

On the other hand, arguably DBMs share such a degree of idealisation with many other computational neuroscience models of an intermediate level of abstraction. In

particular, they are similar in that regard to the closely related Hopfield net and classic connectionist feedforward neural networks. Taking the latter as an example, neural networks have been used in cognitive science (McClelland et al., 2010). What makes them models of processing in the brain is essentially that they capture a specific aspect hypothesised to be important, in a highly idealised manner: that aspect is distributed parallel processing through several layers of neuronal units. What makes DBMs interesting then is that they in a unique way combine such aspects from several models and instantiate several principles debated to play a role in the cortex, as will be elaborated on further in the next chapters.

DBMs learn hidden layers of representations like classic feedforward neural networks, but do so in an unsupervised fashion, being generative models of sensory input. As such models, they can learn V1- and V2-like receptive fields from images (Lee et al., 2008; Saxe et al., 2011), in a similar fashion as sparse coding models (Olshausen & Field, 1996, to be discussed in Section 7.1.1). As in Hopfield nets, low energy states in a DBM serve as (transient) attractor states, but the latter are defined not over observed input patterns that are 'memorised', but rather over internal representations learned from the data. In particular, this means that these internal states can play a role in *perception* rather than just memory. The probabilistic nature of the DBM relates it to other Bayesian approaches in computational neuroscience (Section 1.2), in particular those posing sampling-based computations in the cortex. The probabilistic framework could also be the glue that could bring together models of sensory representation, such as DBMs, and more high-level probabilistic models of cognition (for an example of a model combining both, see Salakhutdinov et al., 2011a).

3.2 In what sense is the DBM a 'model' of the brain?

Ultimately, the issues discussed above lead to the conceptual question of what makes the DBM, or indeed any computational system, a model of the brain. Though potentially fruitful given the current related debate on the status of Bayesian approaches (Section 1.2), an elaborate discussion of these philosophical aspects is beyond the scope of this thesis. We comment on them briefly here (drawing on the general treatments of Frigg & Hartmann, 2008; Webb, 2009).

There are at least two ways how the DBM can be interpreted as a model of cortical processing. First, it can be seen as what it is called an *analogical model*. This view takes the DBM as an information processing system that originated in a non-biological

context (namely, machine learning), and establishes an analogy to cortical processing in terms of computational principles they might have in common (hypothetically), such as analysis by synthesis, by explaining perceptual phenomena on these grounds. Other examples of analogical models would be the billiard ball model of a gas (Frigg & Hartmann, 2008), or cellular automata as an analogy of how global patterns can arise from local rules in biological systems (Webb, 2009). To examine in how far this analogy holds and whether it provides possible insights about the cortex is subject of this thesis.

A second interpretation, more direct and arguably more in line with the way models of cognition are usually thought of, is one of the DBM as an *idealisation* of the cortex, or of analysis by synthesis in the cortex. Idealisation here means "a deliberate simplification of something complicated with the objective of making it more tractable" (Frigg & Hartmann, 2008). The DBM would moreover be a 'Galilean' idealisation, i.e. one that involves deliberate distortions of the properties of the real system for the sake of simplification, here by conflating feedforward and feedback connections in the cortex as simple symmetric connections in the DBM, etc. The DBM would be a rather crude idealisation of the cortex. However, as argued in the last section, it might not be too different from comparable approaches in that regard.

With either interpretation, we would argue that the DBM model can contribute despite its shortcomings considering our current lack of understanding what the important principles underlying cortical processing are, or given a lack of clarity whether such principles can even be formulated across cortical subsystems. The cortex is an extremely complex systems, and it appears particularly difficult to design models that elucidate on and combine several general computational principles rather than just describe individual phenomena, and which attempt to span levels of description from behaviour and high-level perceptual phenomena (such as attention) to neuronal aspects. Such models need to commit to implementation mechanisms that will necessarily deviate from how the cortex solves things, thus making interpreting them more difficult. Ultimately, we see the DBM only as a starting point or point of comparison for more realistic models, which can be approached e.g. by implementing BM type probabilistic inference with spiking neurons (Buesing et al., 2011).

3.3 Interpretation of the latent variables

The hypothesis space of a probabilistic model expresses what kind of knowledge can be inferred from observed data (Section 1.2.2). In a BM, the introduction of hidden units

means that, rather than just expressing constraints between data variables with weights (say, certain pixels in an image tend to be on together), there are now latent variables that can be inferred from the data. With the DBM taken as a biological model, how should these latent variables be interpreted?

In the first chapter, we introduced terminology to clarify aspects of different probabilistic or Bayesian models (Section 1.2.3). In particular, we differentiated between *conceptual* and *instrumental* models, a distinction we will elaborate on here by considering the example of the DBM. The latent variables in a BM appear to be of a rather different nature than those in many other Bayesian models. For example, a Bayesian model might describe the statistical relationship between observed symptoms and the presence or absence of diseases, or a property of the world that needs to be inferred from sensory data, for instance an experimentally controlled variable. In the latter scenario, the Bayesian model might either describe an ideal observer, or, in the case of an internal or psychological model, psychological entities and their relationships possibly employed by the mind. In all cases, by design the variables have a clearly distinguished *meaning* (symptoms, diseases, position of a stimulus, object identity, etc.), possibly even reflecting different functional entities in an external generative process. We called such models *conceptual*. In contrast, in the BM, there is very little built-in assumptions about what a hidden unit represents. Any 'meaning' is imposed onto the unit (in conjunction with its weights) only during learning, determined simply by whatever is useful for the optimisation of the learning criterion (namely the maximisation of the likelihood of the data, Section 2.2). Thus, we suggested the term *instrumental* model here.

In terms of the probabilistic formalism, there is nothing that would inherently distinguish these different models such that the variables of one would be imbued with *a priori* meaning but not those of the other one. Mathematically, all of them can be described as graphical models with latent variables, and they can in principle be learned using the same techniques, such as maximising the likelihood of a subset of observed variables in an essentially unsupervised fashion. The difference of interpretation between the models rather stems solely from the semantics attributed to them by the modeller. In particular, the mindset behind the BM, as is the case with other comparable unsupervised methods, is to *discover* aspects of the data, or to find representations thereof that are useful by some criterion. The assumptions built into the model are more about what structures can in principle be captured, rather than about assigning pre-specified meaning to the involved variables. It is this aspect of these types of unsu-

pervised models that makes them so interesting in our eyes as a data driven, internal models of learning and inference in the brain.

Clarifying how these different types of Bayesian models relate might also be key for mapping high-level conceptual Bayesian models to actual neuronal representations and computations. For example, Kersten et al. (2004) characterise Bayesian models in vision mostly using explicit variables that describe a scene: object classes, illumination direction, viewpoint, etc. They mention models like the Helmholtz machine (related to the BM) when addressing how priors might be learned by the visual system, but do not further discuss how the semantics of these different types of models relate. Similarly, models concerned with how neurons could represent probability distributions (Section 1.2.5) arguably often take the approach that there is a well-defined conceptual, usually low-dimensional variable in the external world (maybe controlled by an experimenter) that is encoded explicitly by the neurons. This approach might tend to neglect the complex computations in the visual system necessary to extract, and possibly learn about, such variables in the first place from sensory input. Conclusions made about explicit probabilistic computations with neurons might then be misleading. We will substantiate this point with a concrete example in the chapter on bistable perception (Section 5.5.1).

Another aspect of BMs that relates to the interpretability of the hidden units is that the latter constitute a *distributed* form of representation (Ackley et al., 1985), meaning that the hidden units together conspire to represent the sensory input and/or that an individual unit on its own cannot necessarily be interpreted to code for any well-defined 'thing' in the sensory input. This is especially true if there is more than one hidden layer. In the first hidden layer, the weights of any hidden units directly connect to the input layer, and, due to the lack of lateral interactions (i.e. the conditional independence assumptions), without higher layers the state of a hidden unit only depends on the data. One can find out about what stimulus activates that unit by inspecting its weights (the 'receptive field'). With multiple hidden layers however, the state of a hidden unit cannot be easily disentangled from the states of the others.

The issue of distributed vs. local types of representation has long been a subject of debate in the connectionist and neural networks fields, and relates to the notion of 'grandmother cells' in the brain. The concern is how a neuron, its activation, and its connections to other neurons represent knowledge. The underlying concepts are not necessarily that clear themselves. For a discussion, see e.g. Bowers (2009) and subse-

quent commentaries.[1] Roughly, a distributed code can use a shared set of components to combinatorially represent many different inputs and share knowledge among different pattern instantiations, and is thus thought to be better at generalisation. A local code on the other hand, where each individual neuron represents a 'concept' of some form, seems more interpretable and maybe more suitable for symbolic representations and high-level cognitive manipulations. Local and distributed codes are not necessarily discrete alternatives, but exist along a continuum (Plaut & McClelland, 2010b). In machine learning, one might encourage more 'interpretable representations' by steering a model to find independent 'causes' of the image, e.g. by building independence assumptions into the model. This is the case in Independent Component Analysis (e.g. Bishop, 2006; Hyvärinen et al., 2009). A similar effect can be obtained in BMs by encouraging sparsity in the hidden representations (Lee et al., 2008; Nair & Hinton, 2009). Sparsity and the nature of the hidden code will play a role in our work on attention (Chapter 6).

Discussing these issues of neural coding further is yet again beyond the scope of this thesis, but we do think they are important. Rather than just being a matter of technicalities of connectionist neural networks, they relate to the crucial question of how knowledge is represented in the brain. For example, Griffiths et al. (2010) contrast their high-level hierarchical Bayesian models of cognition with distributed connectionist networks, arguing that the former are better suited to capture and manipulate 'structured' knowledge. It could also be the case that the brain uses different types of representations in different anatomical systems, for example, in the hippocampus vs. the neocortex (McClelland et al., 1995). What makes models like the DBM interesting is that they bring these representational issues, which are usually associated with neural networks, to probabilistic approaches as well.

Notably, very recent machine learning work (Salakhutdinov et al., 2011b) on learning from few examples (motivated by human capabilities) sees very different types of models combined: a DBM to learn a rich distributed feature space, and a nonparametric tree-structured model to learn a category hierarchy on top of that feature space. Both models are examples of approaches that (in principle at least) can come with little concrete 'meaning' built in *a priori*, rather learning it from the data. They differ however in terms of how knowledge is represented.

[1] Especially, Plaut & McClelland (2010b); Bowers (2010a); Plaut & McClelland (2010a); Bowers (2010b).

3.4 DBM learning: prediction error, dreams, and hierarchical development

In the following, we comment on possible interpretations of the learning mechanisms in the model. At the core of DBM (pre-)training are the RBM learning algorithms, i.e. 'full' BM learning and its RBM specific approximations, CD and PCD (Section 2.2). At first glance, this multitude of algorithms might seem somewhat unwieldy and to lack a clear interpretation from a biological point of view. However, we see some possibly intriguing connections to cortical phenomena. Our work as presented in this thesis does not focus much on the learning itself, but our discussion below suggests there might be promising lines for future research.

As mentioned before, CD, or specifically CD-1, can be understood as being related to reconstruction error driven learning (Bengio & Delalleau, 2009). CD is good at finding a hidden code from which sensory input can be reconstructed. The learning procedure involves inferring a hidden code from the current input, and then making a prediction about the latter from the hidden representation. This thus could capture the idea that the cortex makes predictions about its current sensory input via feedback, as is modelled by approaches such as ART (Grossberg, 1976) or predictive coding (Mumford, 1992; Rao & Ballard, 1999; Friston & Kiebel, 2009). In predictive coding in particular, the hypothesis is that inference is driven by communicating predictions via feedback and resulting *errors* via feedforward connections. Due to the symmetry of the weights, this cannot be implemented in a BM, but CD learning could be seen as a version of prediction driven learning that is matched to this specific idealisation in the BM.[2]

The advantage of CD is that it is fast and involves few computation steps. If interpreted as prediction driven learning, such learning could happen in an online fashion as sensory data comes in, in parallel to perceptual inference itself. What CD appears not be so effective at is learning a good generative model from which data can be sampled freely. Empirically, a RBM trained with CD tends to get stuck in modes of the model distribution; either spurious ones when its state is initialised randomly, or, when the visible units are initialised to a data point, in a mode corresponding to that data point. For example, a RBM trained with CD on handwritten digits and initialised on an image of a 4 will sample 4s for a long time before traversing to other modes. Thus, for learn-

[2] In particular, CD-1 involves computing another hidden code from the reconstructed visible units. Rather than computing the error between original and reconstructed visible units explicitly, the error implicitly is accounted for by computing the difference between unit correlations in positive and negative phases. Perfect reconstruction implies this difference is zero (apart from sampling noise).

ing a generative model, full BM learning or the PCD approximation do a better job, as they decouple the positive and negative phases and allow the model to freely explore the model distribution during the latter, removing spurious modes in the process. This however requires that the model can freely generate 'fantasy' states that are not caused by current sensory input, presumably necessitating some offline time. Taking a BM as cortical model, such a process could possibly be related to dreams. An algorithm like CD could be run during waking to form the basic internal representations, and an algorithm like PCD would refine them during sleep.

As terms such as 'wake-sleep algorithm' in the context of related models suggest (Dayan et al., 1995), the analogy to dreams did not go unnoticed in the machine learning community. Indeed, Crick & Mitchison (1983) took insights from neural memory networks to propose that the function of dreams is exactly to discover and remove ('unlearn') spurious memories, which are an undesired side-effect of learning the actual input patterns. It could be fruitful to reevaluate this idea in the context of the recent development of Deep Learning approaches and algorithms such as PCD. The key here is not so much specifically PCD, which will quite possibly be not the last word in terms of RBM training methods, and it might be rather specific to the RBM model. The underlying 'unlearning' concept however would apply to any model where learning separates into a data driven and a model driven term, in particular, any model formulated in terms of a (general) Boltzmann distribution (Eq. 2.6 on page 21). What PCD does show is that there can be effective approximate algorithms that do not require unfeasibly long sampling to compute the negative phase statistics. Moreover, one of the potential problems (Geoffrey Hinton, personal communication) with identifying the negative phase with sleep or dreaming is that it needs to be performed and alternated with waking phases during each single weight update. Given actual dream and waking periods in animals, the number of updates could thus be too low to allow for effective learning. However, given the various approximations and heuristics already underlying algorithms such as PCD, we would not be surprised if there could be ways around that problem, e.g. by accumulating several positive and negative phase contributions separately in extended waking and sleeping phases. Also, when a separate training algorithm like CD is used during waking, the sleeping phases could focus on removing spurious modes.

One caveat with interpreting CD as learning during waking and PCD-like algorithms with refinement during dreams is that CD only applies to RBMs, not to the more complex DBMs, where essentially PCD is used for learning (plus simplified mean-field inference in the positive phase). It would be interesting to explore whether reconstruc-

tion error driven algorithms can be designed for the DBM, in analogy to CD for RBMs. Alternatively, one could explore the wake/dream relation of the positive and negative phases in DBM learning without appealing to the predictive coding type interpretation of CD.

What about the biological relevance of the iterative layer-wise training procedure, where a DBM is composed by having each subsequent hidden layer learn to generate the activations of the layer below after the latter has been trained? There is good anatomical evidence that the post-natal maturation of higher cortical areas is delayed compared to that of primary sensory ones on a time scale of months (Bourne & Rosa, 2006). Moreso, this maturation appears to occur in a sequential pattern along the cortical hierarchy, especially for the ventral visual stream. The basic principle of layer-wise training that the DBM has in common with other Deep Learning approaches thus appears to be sound from a biological perspective. Again, this could be a promising avenue for further modelling studies. A key aspect of the DBM here would be that even this initial learning is crucially dependent on a higher region making predictions about its respective input patterns, thus suggesting an essential role for feedback connections from the very beginning of the learning process.

3.5 A procedure to decode the internal state

In a machine learning context, a model like a DBM would normally be used either by sampling from it, including the first visible layer, to examine the data the model can produce or 'imagine', or maybe to sample some of the visible units whereas others are observed, filling in missing information; or, the hidden representations inferred would be utilised to some other end, say for feeding them into a classifier to use them to recognise image categories. In the context of this work however, we are concerned with the DBM as a model of perceptual phenomena. In all our applications, we thus feed sensory data to the model by clamping the visible layer to images, and then analyse what is represented in the states of the hidden layers, which define the perceptual state of the model. In particular, we are often interested in aspects of this perceptual content that is *not* matching the visual input, or is not uniquely implied by it: in Chapter 4, we model hallucinatory states in the complete absence of input (blindness). In Chapter 5, we model bistable perceptual shifts caused by ambiguous (but fixed) input; and, in Chapter 6 we examine attentional states that focus on a subset of the visual content in a given image.

One method we use is indeed based on classifying the hidden states (usually those of the topmost DBM layer). We employ a simple neuronal classifier that can itself be treated as part of the BM framework (Hinton et al., 2006). To this end, an additional unit is used that receives the states of the hidden layer as input, but instead of being a standard binary neuron, this "*softmax*" unit can take on m different states, corresponding to a 1-of-m representation of the class variable c. The softmax unit is a generalisation of the binary unit, but it can also be interpreted as a group of m binary units that mutually inhibit each other so that ultimately only one of them is active at a time in a winner-take-all fashion. With w_{ij} being the weight from unit i in this softmax group to hidden unit j, and b_i being its bias, the probability for unit i to be the one that is activated is given as

$$P(c=i) = \frac{e^{b_i + \sum_j w_{ij} h_j}}{\sum_k e^{b_k + \sum_j w_{kj} h_j}}, \tag{3.1}$$

i.e. the unit with the highest total input has the highest probability to switch on. For $m = 2$ classes, the softmax activation probability recovers that of a binary unit (Eq. 2.2 on page 21).[3]

The softmax group is trained in a supervised fashion using labels provided. To this end, it can simply be treated as additional visible units in the respective RBM, with learning rules remaining the same as for binary units. An alternative we did not use is to modify the training cost function to put more emphasis on the correctness of the inferred labels over the normal visible units (discriminative training, Larochelle & Bengio, 2008). In either case, during testing the label is not provided and needs to be inferred by the model instead. Thus the softmax group remains unclamped then, unlike the visible units.

Using the DBM as its own decoder

Classification gives us only limited information about what is represented in the hidden states of the model, and requires class labels to be available in the training data. It also conflates what the DBM has learned to represent with what the classifier has learnt to classify. It would be more informative to have a method that could decode the state of a hidden layer in terms of the sensory data it represents. Recent studies have attempted exactly that for cortical activity, for example by reconstructing movies being watched by subjects or words being listened to, from early visual or auditory regions, respectively

[3] Note that, due to the normalisation, only the relative difference in inputs to the units in the softmax group matters. In particular, this means that the weights and the bias of one of the units can effectively be set to zero by absorbing its parameters into the weights and biases of the other units.

(Nishimoto et al., 2011; Pasley et al., 2012). Pasley et al. (2012) for instance used a simple linear regression model that was trained to predict the auditory spectrogram of the input stimulus from evoked intracranial activities. Whether a similar decoding scheme would work for higher cortical areas is less clear given the several stages of nonlinear processing involved. For the DBM, we can make use of its generative nature and take the same model instance that models a perceptual phenomenon to design a decoder that reconstructs input images from the hidden states. In our applications we find that current representations can differ from hidden layer to hidden layer, thus our method needs to be capable of decoding each individual hidden layer independently.

The most straight-forward case is decoding the states of the very first hidden layer, $\mathbf{h}^{(1)}$. Here, a reconstructed image can be obtained by taking $P(\mathbf{v}|\mathbf{h}^{(1)})$ as a grey-scale image.[4] Again, note that the reconstructed image decoded in this manner might be quite different from the actual input image the visible units are clamped to, e.g. in the case of hallucinations. For decoding any higher hidden layer $\mathbf{h}^{(k)}$, we could take a copy of the DBM, clamp that layer to these states and sample freely from the others (ignoring their actual states in the DBM to be decoded from), producing reconstructed images in the process whenever we sample the visible units. To speed up this process, we instead perform a single deterministic top-down pass, where the activations in each subsequent lower layer are computed using only the layer above as input.[5] This also means that the decoded image is uniquely defined by the specific hidden layer states $\mathbf{h}^{(k)}$, and avoids that the initial state of the decoding model and the dynamics of its lower layers can have an influence on the result, as could be the case if the decoding model were to be sampled from.[6]

We emphasise that our decoding procedure (Figure 3.1) is merely meant as a heuris-

[4] Or by sampling from $P(\mathbf{v}|\mathbf{h}^{(1)})$. Because the visible units are conditionally independent given the first hidden layer and the conditional probability is unimodal, sampling from it will give us no extra information and merely produce noisy versions of the same image.

[5] During the top-down pass, intermediate hidden layers only receive half the normal inputs. Thus, we double the respective weights for the decoding procedure only (cf. the bottom-up initialisation described by Salakhutdinov & Hinton, 2009). Note that, because in general we did not perform fine-tuning of the full DBM model, what makes our model a DBM rather than a deep belief net is the way the stack of RBMs is composed by halving some of the weights. With the doubled weights the decoding procedure actually corresponds to a top-down pass in the corresponding deep belief net. Thus, another way to view the decoding method is that we use a deep belief net based on the DBM for fast decoding. This also means that the decoding might no longer work for a DBM that was further trained after the composition, something we did not test.

[6] Consider an extreme case where the high layer to be decoded from is actually completely disconnected from the lower layers. If we were to decode its states by clamping the latter and conditionally sampling the lower layers, the result could be misleading as the lower layers would still potentially generate a variety of images. In contrast, our procedure uses a deterministic top-down pass, meaning that this degenerate case would be revealed by the fact that the decoded image would always be the same no

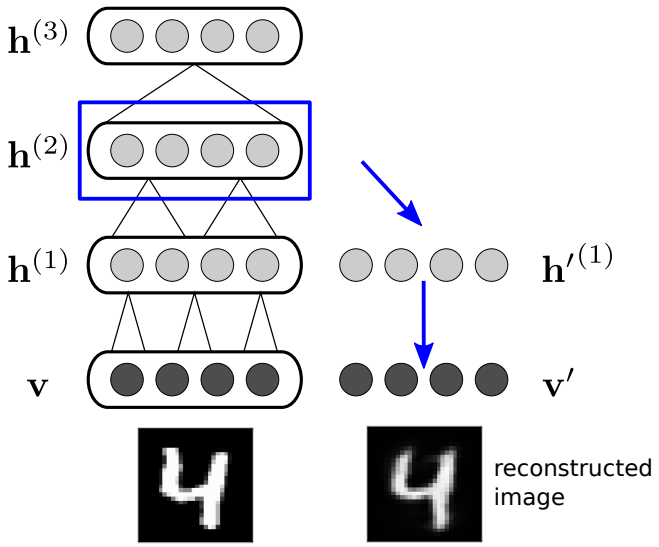


Figure 3.1: During perception, we decode the states of any hidden layer by using the same DBM model as a decoder. Starting from the hidden states of the layer in question, a single deterministic top-down pass is performed to obtain a reconstructed image. Note that this top-down pass ignores the actual states of the other layers in the model instance to be decoded from (see text for explanation).

tic tool for accessing what information is entailed in the internal states of the model. It is not itself part of the perceptual processing being modelled. Alternative decoding schemes could also be explored (e.g. Montavon et al., 2012).

3.6 Model setup

Throughout this work, we used DBMs with similar architectures. Details will be listed in the respective chapters. Generally, the model had three hidden layers with in the order of a 1000 units each, and a visible layer of similar size as determined by the data set used. Unless mentioned otherwise, the weights were constrained as to implement spatially limited receptive fields. Hidden units were arranged in a 2D sheet, and the connections of each unit to the layer below were restricted to a subset of adjacent units. Receptive field sizes increased gradually for subsequently higher layers, and were unrestricted for the topmost layer. We used this configuration to have the model qualitatively reflect receptive field organisation in the cortical hierarchy, and to model specific topographic effects in Chapters 4 and 6.

matter the states of the decoded layer. Ultimately, there is a philosophical issue here concerning what it actually means for a hidden layer to 'represent' something, which we will not address here further.

In terms of notation, we use superscripts to index the different layers and their parameters in the DBM, starting with 0 for the visible layer. For instance, $\mathbf{x}^{(0)}$ and $\mathbf{W}^{(0)}$ refer to the states of the visible layer and the weights between the visible layer and the first hidden layer, respectively.

3.6.1 Training procedures and parameters

For training, we used either CD-1 or PCD for layer-wise pre-training, where each pair of adjacent layers forms a RBM, and no fine-tuning of the full DBM. It should be noted that learning in BM type models can be somewhat of a (black?) art, as there are various meta-parameters that affect the outcome, such as the number and sizes of the hidden layers and the learning rate, and only heuristic recipes for setting them (Hinton, 2010b). There are also additional tricks like using momentum (meaning that the weight updates are smoothed by averaging the current gradient with recent ones) and weight decay (op. cit.). For our work, we based our initial implementation on publicly available deep belief net code (Hinton & Salakhutdinov, 2006), and kept meta-parameters in default ranges unless a change was necessary (e.g. when implementing PCD over CD). In general, our study was concerned with simulating qualitative phenomena rather than with breaking performance benchmarks or matching quantitative experimental data. To achieve the results to be presented, we found that little parameter fiddling was necessary once a model setup worked in principle.

Below we list the parameters values used throughout this thesis for training the RBM components of the DBM. Training aspects that varied from case to case (model sizes and data sets used) will be described in the respective result chapters. The general procedures and parameter ranges adhered to the ‘practical guide’ to training RBMs of Hinton (2010b), to which the reader should refer to for further elaboration. The underlying concepts were explained in Chapter 2, Section 2.2 (see Figure 2.3 on page 34 for a graphical overview). CD-1 or PCD were used depending on what was found to work better for the data set in question. Depending on the training variant, some training parameters differed: it was important to adjust the order of magnitude of the learning rate and the momentum meta-parameter to make the two methods work. Other differences had only a smaller impact, and were in part incidental aspects of exploring different training setups at the time.

All training data sets had 60,000 images and were split into minibatches of 100 images each. At the beginning of training, the weights were initialised randomly using

values drawn from a zero-mean Gaussian with a standard deviation of 0.1. All biases were initialised to -4 to encourage sparsity (to be elaborated on in Section 6.4 in the chapter on attention). Training then proceeded over 30 epochs (i.e. iterations through the training data). Weights were updated using weight-decay with a weight-cost of 0.0002. If CD-1 was used, then the learning rate was kept fixed at 0.1; initial momentum was 0.5 for the first six epochs and then changed to 0.9. In the case of PCD on the other hand, the learning rate was initialised to 0.005 and then decreased linearly to 0 over the course of training; momentum for PCD was kept at 0.5 throughout; the Markov chain was run for 5 steps in the negative phase, with a number of fantasy particles matching the size of a minibatch (i.e., 100); and, activation probabilities were used for the visible states in the negative phase rather than sampled binary states. Lastly, to train the next RBM in the DBM, activation probabilities of the current topmost hidden layer were used as data.

Chapter 4

The synthesis of hallucinations in Charles Bonnet syndrome

Visual hallucinations can offer fascinating insights into the mechanisms underlying perceptual processing and the generation of visual experience in the brain. Of particular interest is a pathology known as Charles Bonnet syndrome (CBS), for two reasons. First, hallucinations in CBS can be very complex in the sense that they entail vivid, life-like, and elaborate imagery of objects, people, animals, or whole visual scenes. Second, the primary cause of CBS is loss of vision due to eye diseases, with no clear pathology in the brain itself and no necessary impairment to mental health other than the hallucinations. De-afferentation of the visual system and sensory deprivation thus seem to be the important factors in the development of CBS, and comparisons have been made to phantom limb phenomena. Unlike for example in the case of schizophrenia, most often accompanied by auditory hallucinations (Mueser et al., 1990), in CBS there thus does not seem to be a more pervasive malfunction of the cognitive system, but rather some form of over-compensation or maladaptation of the relatively healthy brain to the lack of sensory stimulation.

From a theoretical perspective, there has been an attempt to unify complex visual hallucinations in various pathologies in a single qualitative model (Collerton et al., 2005), but many argue that the underlying causal mechanisms are too varied to do so (ffytche, 2005; Morrison & David, 2005; Spencer & McCarley, 2005). At the same time, the fact that hallucinations can occur in many different circumstances speaks to them relating to essential aspects of perceptual inference in the brain. In particular, theoretical explanations that pose that perception inherently involves some form of active synthesis of internal representations might be well positioned to shed light on the generation of

spontaneous imagery in hallucinations, which can occur even in cases such as CBS where there seems to be little defect in the visual system other than at the input stage. Therefore, two key questions arise here: what do complex hallucinations tell us about general principles of cortical inference, and what are the mechanisms triggering CBS in particular?

The purpose of this chapter is hence threefold. First, to gain theoretical insights into important principles of cortical inference, we employ the deep Boltzmann machine (DBM) as model system which is based on such (hypothetical) principles. Second, to examine concrete causal mechanisms for CBS, we model homeostatic regulation of neuronal firing activity, elucidating on various aspects of CBS. Moreover, to examine a potential role of the neuromodulator acetylcholine, we introduce a novel model of its action as mediating the balance of feedforward and feedback processing in the cortical hierarchy. And third, with our results we aim to demonstrate the relevance of the DBM as model of cortical processing. An early version of the presented work has been published (Reichert et al., 2010).[1]

In the next section, we give a more detailed overview over CBS, discuss the relationship between hallucinations and cortical inference, and introduce neuronal homeostasis as a potential mechanism. In Section 4.2, we describe how hallucinations and homeostasis can be modelled with the DBM, including details of the simulation experiments performed. In Section 4.3, the results are presented, covering in particular: a possible functional role of homeostasis in the context of a generative model in the cortex; how hallucinations are caused by visual degradation in CBS; how aspects of the visual degradation (such as localisation in the visual field) correlate with hallucinatory content; how lesions or suppression of activity at different stages in the cortical hierarchy influence hallucinations; and, the role of acetylcholine. Lastly, in Section 4.4, we summarise our results, and discuss remaining issues with CBS and what the syndrome implies for cortical processing and the generation of visual experience.

4.1 Charles Bonnet syndrome

CBS is characterised by complex recurring visual hallucinations in people who suffer from visual impairment but no other psychological condition or hallucinations in other

[1] Compared to the earlier publication, the key changes are the utilisation of sparse representations and the following additions: the MNIST data set; large objects in the shapes data set to elucidate on "Lilliputian" hallucinations; the sensory impoverishment experiment; modified and additional ACh experiments; and more extensive analysis and discussion throughout.

modalities (Schultz & Melzack, 1991; Teunisse et al., 1996; Santhouse et al., 2000; Menon et al., 2003). In particular, patients generally gain insight into the unreality of their experiences. The phenomenology of CBS is multifarious, with the nature and content of hallucinatory episodes as well as the conditions favouring their occurrence varying from patient to patient or episode to episode. Common themes are the vividness and richness of detail of the hallucinations, the elaborate content often entailing images of people or animals (though often of a bizarre nature–figures in elaborate costumes, fantastic creatures, extreme colours, etc.), as well as possibly common triggers, such as being in a state of drowsiness and low arousal. Episodes can last from seconds to hours, and hallucinations can reoccur over periods lasting from days to years.

The eponym CBS itself is somewhat ambiguous or even controversial (ffytche & Howard, 1999; Menon et al., 2003; Plummer et al., 2007; ffytche, 2007). Some authors put the emphasis on complex hallucinations in visually impaired but psychologically normal people, where the visual pathology can be anywhere in the visual system from the retina to cortex; others define CBS to be necessarily related to eye diseases only. Similarly, the delineation of the term 'complex', and whether CBS should include complex hallucinations only, appears to be not fully clear. On one end are simple or elementary hallucinations consisting of flashes, dots, amorphous shapes, etc., while on the other are fully formed objects or object parts like animals, people, and faces (Menon et al., 2003; Collerton et al., 2005). Somewhere in between are geometric patterns ('roadmaps', brickwork, grids, and so forth). Some authors include the latter in CBS (Burke, 2002; ffytche, 2007). It should be noted that simple hallucinations are actually more common in visually impaired patients than complex ones, with a prevalence of about 50% vs. about 15%, respectively (Menon et al., 2003). Both types can occur in individual subjects, possibly with a tendency to progress from simple to complex over time (Menon et al., 2003).

For this modelling study, we identify the following key aspects of CBS we aim to capture and elucidate on. First, we take the common definition of *hallucinations* as compelling perceptual experiences in the absence of external stimuli. They are to be contrasted (Menon et al., 2003; Collerton et al., 2005) to illusions as misperceptions concerning an actual external stimulus, as well as to mental imagery. Unlike hallucinations, the latter is under complete volitional control, lacks perceptual vividness (it appears to be 'in the mind's eye' rather than in the world), and might also have a different neurobiological substrate (ffytche, 2007).

Second, in the context of CBS we are interested in hallucinations that are *perceptu-*

ally rich in the sense that the experience is similar to that of actual seeing. Presumably, this implies that the representations instantiated in the neuronal activity patterns share significant commonalities in both seeing and hallucinating, though this requires further elaboration.

Third, we consider hallucinations on the *complex* end of the spectrum, i.e. objects, people, and so forth. As we currently lack good generative models of realistic images (biological or otherwise[2]), the model we employ still relies on relatively simple binary images. However, it attempts to capture at least some aspects of how complex, object-based hallucinations might be created in the brain. For example, the content of complex hallucinations presumably cannot be accounted for by appealing to anatomical organisational properties of lower visual areas, which Burke (2002) suggested for simpler hallucinations of geometric patterns in CBS (referring to anatomical "stripes" in V2 etc.). Our model relies on distributed, high-dimensional, hierarchical representations that go beyond local low-level visual features (e.g. V1 like edge detectors). The representations are learnt and reflect structure in sensory data beyond local correlations.

Fourth, with regards to the issue of whether CBS should refer to hallucinations in the context of eye diseases only, our model is meant as a model of processing in the cortical hierarchy, and due to the level of abstraction we only require that *visual input is lost* somewhere at a preceding stage and do not differentiate further. We do however address the distinct roles of cortical areas within the hierarchy in Section 4.3.5.

CBS is a complex phenomenon with manifold symptoms and little data beyond clinical case reports and case series. The aim of our computational model is thus to qualitatively elucidate on possible underlying mechanisms, to demonstrate how several common aspects of CBS could be explained, and to gain some potential insights into the nature of cortical inference.

4.1.1 Hallucinations, cortical inference, and acetylcholine

The occurrence of complex visual hallucinations in various pathologies (Manford & Andermann, 1998; Collerton et al., 2005) as well as the imagery we all experience in dreams show that the brain is capable of synthesising rich, consistent internal perceptual states even in the absence of, or in contradiction to, external stimuli. It seems natural to consider hallucinations in the context of theoretical accounts of perception

[2]Naturally we are referring here to generative models from neuroscience or artificial intelligence which can be inverted to make perceptual inferences over images, rather than *purely* generative algorithms from computer graphics.

that attribute an important functional role to the synthesis of internal representations in normal perception, not just in pathological conditions. In particular, the notion of perception entailing 'analysis by synthesis' has been elaborated on in the framework of hierarchical Bayesian models of vision (Yuille & Kersten, 2006). The idea is that ambiguous sensory signals inform initial hypotheses about what is in an image in a bottom-up fashion (from low-level image features to high-level concepts, like objects and faces). These hypotheses are then made concrete in a synthesis stage that tests a hypothesis against the image (or low-level representation thereof) by making top-down predictions using a generative process.

As we have discussed in Section 1.2, the term Bayesian is rather ambiguous. Here, the relevant aspect is that Bayesian models might offer a way of formalising notions of 'bottom-up' processing driven by sensory input, and internally generated, 'top-down' processing conveying prior expectations and more high-level learned concepts. More concretely, hierarchical Bayesian models, rather than just defining some 'ideal' observer (Yuille & Kersten, 2006), could offer ways of interpreting hierarchical cortical processing (Lee & Mumford, 2003). Top-down processing then corresponds to information flow from higher areas to lower areas, and inference is implemented via recurrent interactions between cortical regions. The role of feedback to lower areas would either be to evaluate the top-down predictions (predictive coding), communicating residual errors in feedforward signals, or to inform lower level hypotheses about their consistency with high-level ones, or a mixture of both (as discussed by Lee & Mumford, 2003). An imbalance of, or erroneous interaction between, such bottom-up and top-down processing is then a possible cause underlying hallucinations (Collerton et al., 2005; Corlett et al., 2009). Bayesian models have thus been suggested to shed light on hallucinations (Yu & Dayan, 2002; Friston, 2005; Corlett et al., 2009), although this has to our knowledge not been explored concretely in a computational model.

It should be noted that the general notion of higher cortical areas conveying predictions to lower areas itself does not necessarily imply a Bayesian model, in the sense of one that explicitly deals with uncertainty.[3] Some computational models implementing this notion thus lack the probabilistic aspect (e.g. Adaptive Resonance Theory, Grossberg, 1976, which has also been related to hallucinations, Grossberg, 2000), or do not emphasise it (the predictive coding model by Rao & Ballard, 1999). On the other hand, the brain's treatment of uncertainty could itself be key for determining the effective

[3] And in turn, a Bayesian model does not necessitate a generative, analysis by synthesis component (Yuille & Kersten, 2006).

balance between bottom-up and top-down: this is because in a Bayesian model, the uncertainty associated with either the bottom-up sensory information or the top-down priors will determine how much either will influence the final result of inference. In the model of Yu & Dayan (2002), the authors assign a concrete biological mechanism to represent the uncertainty of the prior, hypothesising that the neuromodulator acetylcholine could play this role. They thus refer to its relevance in some hallucinatory pathologies as evidence, where deficient acetylcholine, corresponding to an over-emphasis of top-down information in Yu & Dayan's account, could lead to hallucinations (Perry & Perry, 1995; Manford & Andermann, 1998; Collerton et al., 2005).

As Yu & Dayan (2002) state, a shortcoming of concrete Bayesian models such as theirs is that they are often formulated over very simple, low-dimensional, non-hierarchical variables. It is not clear how their treatment of priors and uncertainty translates to models that deal with high-dimensional problems (like images) in a biologically plausible manner. This is what we need to address if we hope to develop a computational model of CBS, and in this context we will introduce a novel model of the action of acetylcholine in similar spirit to the framework of Yu & Dayan (Section 4.3.6).

4.1.2 Neuronal homeostasis as causal mechanism

While hallucinations in general might relate to an imbalance of bottom-up and top-down in the cortex, the causes behind specifically CBS and the involved mechanisms are poorly understood (for discussion, see Schultz & Melzack, 1991; Manford & Andermann, 1998; Menon et al., 2003; Plummer et al., 2007). Evidence from CBS and other pathologies suggests that an intact visual association cortex is necessary as well as sufficient for complex visual hallucinations to occur (e.g. Manford & Andermann, 1998). For example, lesions to visual cortex can cause hallucinations, but only if they are localised to earlier areas and do not encompass the higher association cortex. One of the insights emerging from the debate is that the pathology in CBS appears to entail primarily a loss of input at stages prior to association cortex. In contrast, for example for hallucinations accompanying epilepsy there is thought to be an irritative process that directly stimulates association cortices.

How deficient input in CBS leads to the emergence of hallucinations is unclear. Classic psychological theories suggest that the lack of input somehow 'releases' or disinhibits perceptual representations in visual association cortex. This somewhat vague notion has been made more concrete by taking neuroscientific evidence into account

which shows that cortex deafferentiated from input becomes hyper-excitable and generates increased spontaneous activity. As Burke (2002) argues (also Plummer et al., 2007), changes to neuronal excitability as a consequence of decreased presynaptic input, based on for example synaptic modifications, could thus underlie the emergence of neuronal activity which establishes hallucinatory perception in CBS.

Such adaptive changes of neuronal excitability have been studied extensively over the last two decades in experimental and theoretical work on *homeostatic* plasticity (see Desai, 2003 for review; also Marder & Goaillard, 2006; Turrigiano, 2008). Rather than deeming them artifacts or epiphenomena, such changes have been attributed important physiological functions, allowing neurons to self-regulate their excitability to keep their firing rate around a fixed set-point. Homeostatic regulation is thought to stabilise activity in neuronal populations and to keep firing within the neurons' dynamic range, compensating for ongoing changes to neuronal input either due to Hebbian learning, or due to developmental alterations of the number of synapses, connectivity patterns, etc.

A neuron might track its current activity level by measuring its internal calcium levels, and several cellular mechanisms have been identified that could then implement homeostatic adaptation. Among them is 'synaptic scaling', a change to synaptic efficacy that is thought to affect all synapses in a neuron together, keeping their relative strengths intact. Alternatively, the intrinsic excitability of a neuron can be regulated by changing the distribution of ion channels in the membrane. Both mechanisms have been observed experimentally, dynamically regulating neuronal firing rate over a timespan from hours to days (Turrigiano & Nelson, 2000) in compensation for external manipulations to activity levels–in particular, in response to an activity decrease caused by sensory deprivation.

Hence, with visual input degraded due to eye disease or other defects in the visual pathways, homeostatic *over*compensation is a strong contender to be the neuronal cause underlying the emergence of hallucinations in CBS. This is the mechanism we explore in our computational model.

4.2 The DBM model of hallucinations

To address CBS, we need to work towards computational models that can capture its key properties as identified earlier. Such a model should be able to internally synthesise rich representations of image content, such as objects, even in the absence of (corresponding) sensory input.

The DBM model was introduced in detail in Chapter 2. We argue that the DBM is promising as a model of hallucinations because it is a generative model that learns to synthesise representations of sensory data. A DBM can be seen as an instance of a hierarchical Bayesian model, and thus could capture the intuition of bottom-up and top-down processing in the cortex reflecting the interaction between sensory information and internal priors. An imbalance of such processing then can be seen as a cause for hallucinations to emerge. At the same time, the DBM is also as a simple neural network, thus enabling us to explore concrete neural mechanisms possibly underlying CBS. Unlike the related Hopfield network, which itself has been used to model hallucinatory 'memories' in schizophrenia (Ruppin et al., 1996), it does not just memorise given input patterns, but rather learns *internal* representations of input images. This makes the DBM a more concrete model of *perception* rather than just memory. The 'deep' organisation of the DBM into hierarchical layers as well as the image based representations will allow us to make some concrete connections to the visual cortex.

4.2.1 Homeostasis in a DBM

We model CBS as resulting from homeostatic regulation of neuronal excitability in response to degraded visual input. We use DBMs that have learned to represent images, having trained them on either of two simple data sets. We then simulate the visual impairment by using empty or corrupted input instead of the original data, and have the model perform inference over them. The change in sensory input could lead to changes in the activation levels of the model's neuronal units. To model homeostatic mechanisms, we allow the neurons to adapt their excitability in response.

As discussed above, homeostatic plasticity can be described as a neuron adapting its excitability to match its current average firing rate (as measured over hours or days) to a fixed set-point (Turrigiano, 2008), and there are several cellular and synaptic processes making this possible. Here, for simplicity we model a single basic mechanism, namely an iterative adaptation of each neuron's intrinsic excitability. With target activity p_i and current average activity a_i, neuron i in the DBM should become either more or less excitable according to the difference $p_i - a_i$. Its bias parameter b_i is thus iteratively incremented by

$$\Delta b_i = \eta(p_i - a_i), \tag{4.1}$$

where η is a constant parametrising the rate of adaptation. Such an adaptation of the bias has the effect of shifting the activation function of the unit, i.e. the probability for it

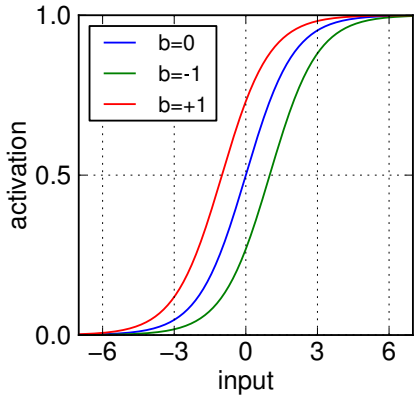


Figure 4.1: The activation probability (given by the logistic function) of a neuron shifts depending on the value of the bias parameter b.

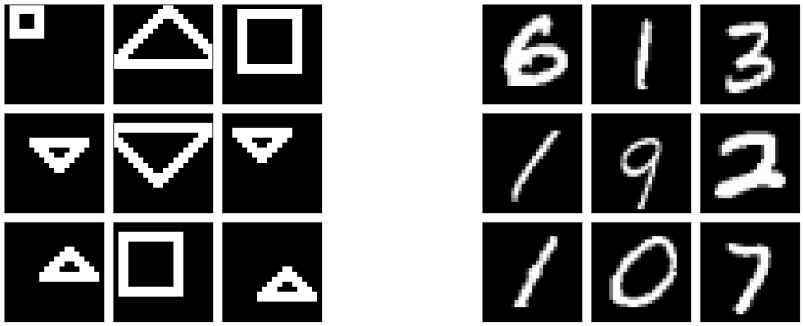
to switch on, rendering it more or less excitable for a given amount of input (Figure 4.1; cf. Figure 3a in the paper by Desai, 2003, on homeostatic plasticity).

To define the target activity p_i for each neuron, we simply assume that the average activity of a unit during inference over the training data (after training) defines the normal, 'healthy' level of activity for the representations learned. An alternative possibility would be to impose a fixed level in advance, and use the homeostatic mechanism during training itself to encourage units to attain it. This corresponds to a regularisation that has indeed been used for related machine learning models, e.g. to enforce sparsity in the representations (Lee et al., 2008; Nair & Hinton, 2009). We report here results without using this mechanism in training itself.[4] We did however try the latter and obtained similar results. Thus, what mattered here is only that whatever activity levels had been assumed during training were restored, regardless of whether these levels were originally confined to a certain regime.

4.2.2 Methods and model setup

We used two training data sets to explore different aspects of CBS (Figure 4.2). The first is a custom set of binary images containing toy shapes of various sizes at various positions. This shapes data set allowed us to examine issues related to the localisation of visual impairment, and due to its simplicity the perceptual content of the corresponding hallucinations is straightforward to analyse by directly comparing it to training images. The second data set is MNIST, which contains images of handwritten digits and is

[4] We did use standard weight decay during training however, which could be seen as another type of homeostatic mechanism akin to synaptic scaling.



(a) Shapes. (b) MNIST.

Figure 4.2: Examples from the training data sets (see main text for details). *(a)*: a custom data sets of simple shapes at various positions. *(b)*: the MNIST data set of handwritten digits, a standard benchmark in machine learning.

a standard benchmark used in machine learning. The advantage of MNIST is that it contains objects that, if still simple, arguably have some more interesting structure. With such kinds of data it has been shown that DBMs can learn representations that generalise to unseen instances of the data, not just in terms of classification performance but also in terms of the data they generate themselves (Eslami et al., 2012). This in particular demonstrates that learning does not simply correspond to memorising training images.

For both data sets, the employed DBMs had three layers of hidden units, using topographic restricted receptive fields (Section 3.6). The biases of the units were initialised to negative values before training to encourage sparse representations. In particular, this lead to a breaking of symmetry between on and off states, and to input degradation (which models visual impairment) generally having the effect of decreasing neuronal activity (this role of sparse representations will be elaborated on in the chapter on attention, Section 6.4[5]). Consequently, homeostatic regulation then would have to recover firing rates by increasing the excitability of the units. This matches the findings that cortical neurons become 'hyper-excitable' under sensory deprivation (as reviewed e.g. by Burke, 2002). Other than the sign of the activity changes, overall results as reported in this chapter did not however depend on representations being sparse.

For MNIST, the visible layer had 28×28 units corresponding to the size of the images in pixels, and 28×28, 28×28, and 43×43 units in the three hidden layers, from lowest to highest, respectively. Receptive field sizes were 7×7, 14×14, and 28×28.

[5] Briefly, by encouraging units to be off most of the time, they learn representations where they signal the presence of specific content in an image by switching on. Thus, removing input tends to make them turn off.

The model was trained layer-wise for 30 epochs (i.e. iterations through the training data) in each layer, using 5-step PCD. The training set contained 60,000 images, 6,000 per digit category (0 to 9). For the shapes data set, the visible layer had 20×20 units and the hidden layers 26×26 units each, with receptive field sizes 7×7, 13×13, 26×16. Here, layer-wise training consisted of 30 epochs of CD-1. The training set again had 60,000 images in total, from six categories (squares, triangles in two orientations, all in two different sizes). It should be noted that, due the limited variability in the shapes data set, all possible image instances were covered by the training set. Hence, only the MNIST data set is suitable to test the generalisation performance of the model. Lastly, for neither MNIST nor the shapes data set were the models trained further after the layer-wise pre-training. See Section 3.6.1 for further details on the training parameters used for CD-1 and PCD.

To measure the preferred activity p_i for each hidden neuron, we averaged its activation over all training data (after learning), with one trial per input image consisting of 50 sampling cycles. Here and elsewhere, the hidden states were generally initialised to zero at the start of a trial. Similarly, to measure the current average activation a_i during homeostatic adaptation, activities were measured over 50 cycles in 100 trials per iteration. Depending on the experiment in question, the visible units were set to a different image for each trial or remained blank (when modelling complete blindness). The adaptation rate η was set to 0.1 and 0.04 for models trained on shapes or MNIST, respectively.[6]

To analyse the perceptual state of the model, we decoded the states of the hidden layers as described in Section 3.5, obtaining a reconstructed image for each layer at each sampling step. To evaluate the internal representations w.r.t. their possibly hallucinatory content, we analysed whether the decoded images corresponded to the kind of objects the models had learned about in training, using the topmost hidden layer's states after 50 sampling cycles for quantitative analysis. For the shapes data set, we employed a simple template matching procedure, matching the image to the shape templates used in training by convolving the former with the latter.[7] The maximum value of the resulting 2D vector was taken as quantitative measure for the correspondence, termed the 'hallucination quality', where a perfect match corresponded to a hallucination quality of 1. For the more varied MNIST data set, there are no fixed templates, nor

[6]The MNIST model was found to effectively adapt faster, thus a lower rate was chosen to obtain a higher temporal resolution. In terms of overall results, the precise value of the rate did not matter too much.

[7]Each image had its mean subtracted and was then l^2 normalised.

do generated images necessarily match instances from the training set (which is the point of having a model that can generalise, as mentioned above). To obtain a measure of hallucination quality, we classified the decoded image as belonging to any of the digit categories, using the confidence of the classifier as a measure of the image's quality. Specifically, we used an instance of the DBM model itself (not affected by homeostasis) with a classification unit attached (Section 3.5). Taking the maximum of the posterior over the digit categories again yielded a measure with maximum value 1. Inspecting the generated image and resulting posterior values, we also confirmed that for images that did not look like well-defined MNIST digits, classification scores computed in this manner tended to be lower.

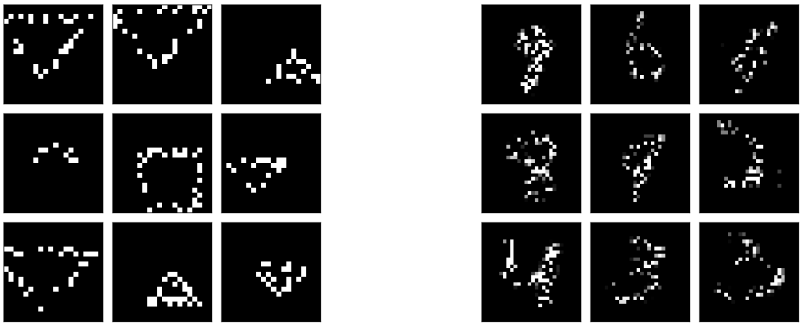
4.3 Experiments

The hypothesis we explored is that homeostatic regulation of neuronal firing rate in response to sensory deprivation underlies the emergence of hallucinations in CBS. The possibility for synthesis of internal representations is explained by the cortex implementing a generative model of sensory input. As a first step, we aimed to demonstrate that the homeostasis mechanism as implemented in the model can actually be beneficial in this context.

4.3.1 Robust analysis by synthesis due to homeostasis

Homeostatic adaptation is thought to stabilise neuronal activity in the face of circuitry changes due to learning or development. In the following, we show how it could be helpful in particular for a model that implements perceptual inference by synthesising internal representations, by making the learned representations robust against exactly the sort of visual degradation that ultimately causes CBS. To this end, we had the model (trained on either the shapes or MNIST data sets) perform inference over heavily corrupted versions of the images (Figure 4.3). The latter were created by taking images from the data sets (digit instances not seen in training in the case of MNIST) and setting 65% of the pixels to black.

Degrading the input in this manner lead to profound activity changes in the neurons, which the model was then allowed to compensate for by employing homeostatic adaptation. Figure 4.4 shows how activity levels changed under input degradation and subsequent adaptation, plotted either against the number of preceding iterations or the



(a) Corrupted shapes. (b) Corrupted MNIST.

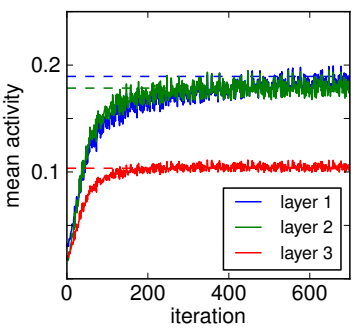
Figure 4.3: Examples of the corrupted images used to test whether homeostatic adaptation can be beneficial by restoring internal representations.

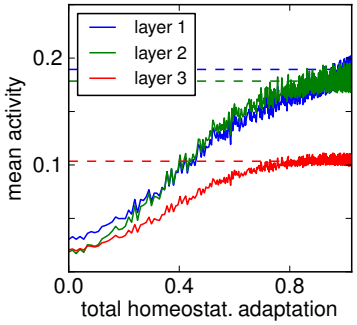
total shift of the bias parameter so far (averaged over all units). For all three hidden layers, initial activities were lower when compared to normal levels. Homeostatic adaptation then led to a gradual restoration to the original values.

Importantly, this recuperation of activity levels corresponded to a restored capability of the model's internal representations to capture the underlying objects in the images. We decoded the hidden states of the top layer and classified the resulting reconstructed images using a classifier trained on the original data sets. Input degradation initially lead to a sharp drop in performance in classifying the corrupted images (Figure 4.4(a, d)). However, homeostatic adaptation lead to a significant improvement of classification, reaching a performance that was close to the one achieved on the decoded representations inferred from uncorrupted images.[8]

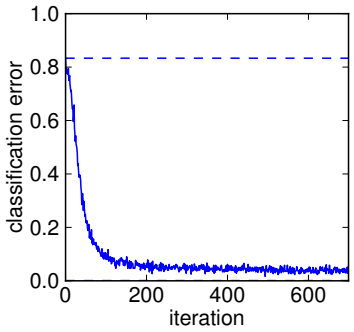
Hence, the homeostatic mechanism as defined by Eq. 4.1 can be sufficient to restore the representations inferred over sensory input as to be suitable for classification. This is despite the fact that it only attempts to match the average activations, i.e. first order statistics of the inferred posteriors averaged over all input images, rather than the full distribution learned in training, and only does so by adapting the bias parameters. Thus, homeostatic adaptation could offer a simple local neuronal mechanism that serves to

[8] As in the context of measuring the hallucination quality for MNIST (described in Section 4.2.2), another instance of the DBM was used as the classifier. For MNIST, on decoded representations inferred from uncorrupted images, the error rate was 8.8% (compared to 7.0% when classification was performed directly on the hidden states rather than reconstructed images). It should be noted that the aim of our work was not achieving high classification performance, hence we did not train the full model, fine-tune the hyper-parameters, nor necessarily implement classification in an ideal fashion. Classification is merely used to analyse the quality of the internal representations. The reported score for MNIST is hence lower than the state of the art, the latter being around 1% for this type of model (e.g. Salakhutdinov & Hinton, 2007).

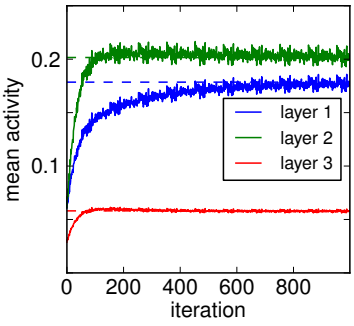


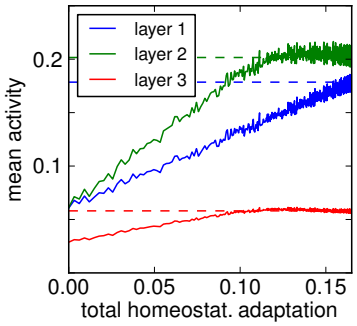


(a) Corrupt shapes activity (vs. iterations or bias adapt.).

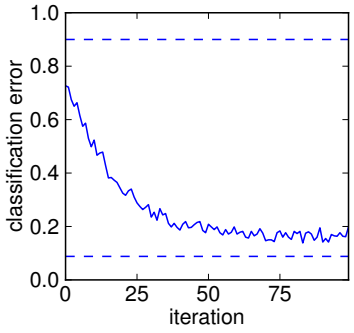


(b) Corrupt shapes classification.





(c) Corrupt MNIST activity (vs. iterations or bias adapt.).



(d) Corrupt MNIST classification.

Figure 4.4: With corrupted sensory input, homeostatic adaptation led to a restoration of the internal representations. *(a)*: average activity levels in each of the three hidden layers over the course of homeostatic adaptation. The left figure shows activity levels plotted against the number of iterations so far, the right figure against the total homeostatic adaptation in the neuronal bias parameters (absolute differences between current bias values and initial values, averaged over all units). Dashed lines correspond to normal activity levels for each layer with uncorrupted input. Activities initially dropped profoundly as input was corrupted, but then recovered as the neurons adapted. *(b)*: classification error using the internal representations to classify the corrupted input (see text for details). Top dashed line is chance, bottom one is performance on uncorrupted input (here, for the simple shapes data set, the error is very close to zero, hence the corresponding line is drawn on top of the x-axis). Over the course of adaptation, internal representations are restored as well, allowing for classification performance close to its original level. *(c)+(d)*: analogous to (a) and (b) but for a model doing inference over corrupted MNIST images.

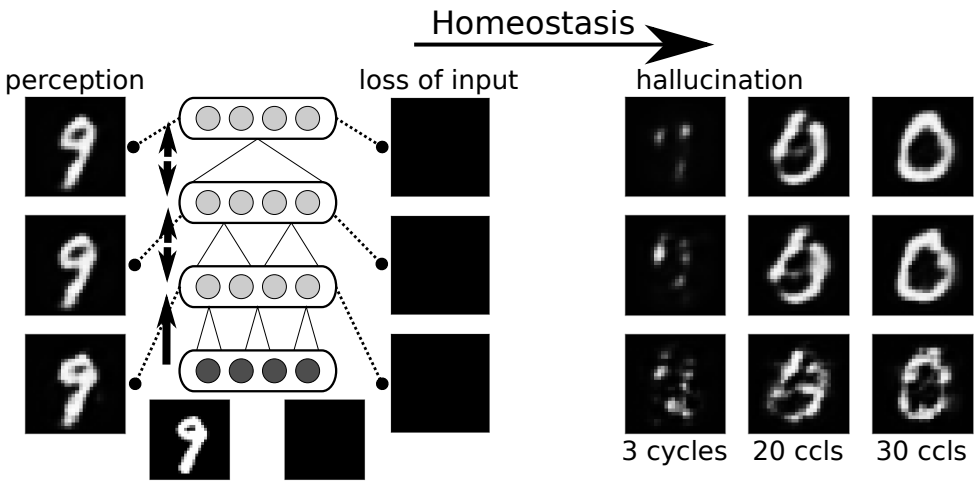


Figure 4.5: Overview of the basic CBS experiment. A model has been trained on simple images (here, MNIST digits). Initially, decoded internal representations correspond to what is given as input in the visible layer. To model visual impairment or blindness, sensory input is then removed, eliciting internal representations devoid of content. Subsequent homeostatic adaptation of neuronal excitability leads to spontaneous hallucinatory representations emerging (right-hand side images are decoded from the hidden layers, receiving no sensory input, 3, 20, or 30 sampling cycles after initialisation).

make learned representations robust for example against degradation of sensory input. It does not rely on further learning (in the sense of parameter changes that incorporate incoming sensory data), intricate synaptic changes, or network wide measurements. Rather, each neuron only needs to remember its average activity level and regulate its intrinsic excitability accordingly. However, as we will see in the following, this stabilisation of perceptual representations can be detrimental, ultimately decoupling internal representations from a further degraded sensory input, causing hallucinations.

4.3.2 Emergence of hallucinations

To model more profound visual impairment or blindness, we then repeated the above experiment but with the visible units permanently clamped to completely empty input. As before, the model had initially been trained on images from either of the two data sets. Then the input was exchanged, and homeostatic adaptation was allowed to take place. Any emergence of meaningful internal representations in the absence of input would correspond to hallucinations. See Figure 4.5 for an overview of the CBS experiment.

Before presenting the results, we should briefly comment on how the binary input images are to be interpreted so that presenting a blank image corresponds to ‘taking

the input away', i.e. blindness. After all, if these were grey-scale images, then seeing a black image would not be the same as not seeing altogether. Rather, the binary images are to be understood as proxies of images already having been encoded in neuronal activity at an early stage of visual processing (e.g. primary visual cortex), which we do not model here for simplicity (but note that the latter exactly corresponds to the situation in experiments we report on later where we model loss of vision in higher stages of the hierarchy).

Figures 4.6(a, b) and 4.7(a, b) show the activity changes resulting from visual impairment and subsequent adaptation, for two models trained on either shapes or MNIST, respectively. Again we found an initial drop of activity that was subsequently fully compensated for, at least on average over each hidden layer, by the shift of the intrinsic excitability of the neurons.

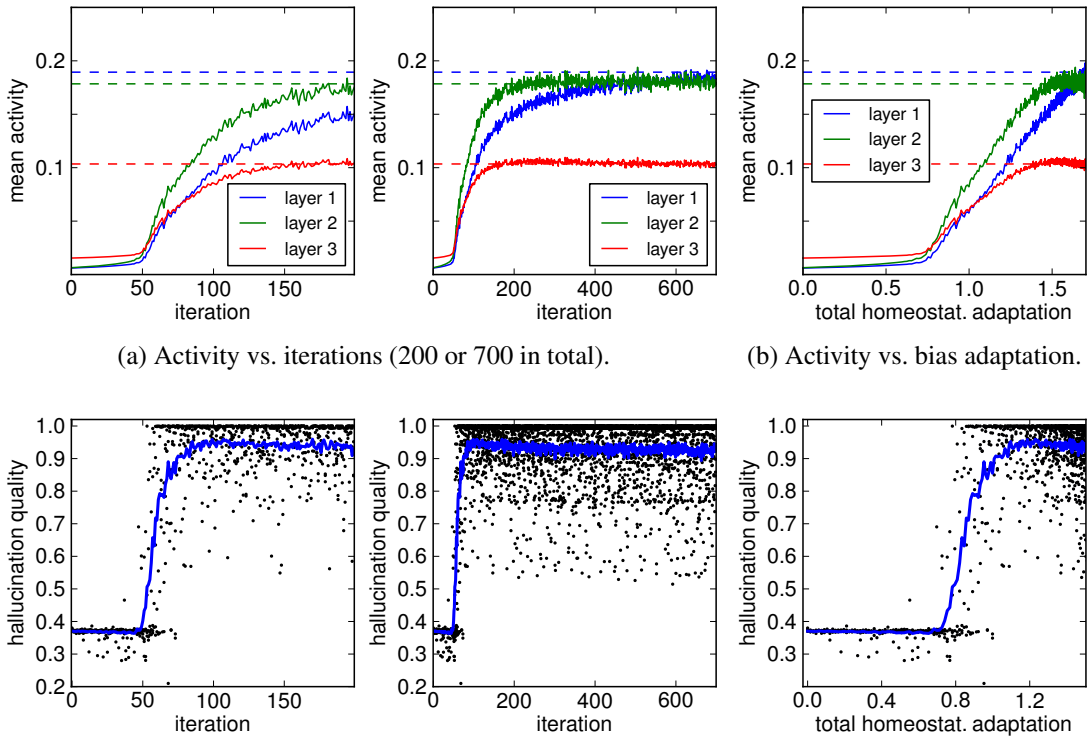
The question then was what the nature of the internal representations was that allowed for a restoration of activity levels. After all, the purely local adaptation of each neuron might have recovered individual preferred firing rates on the basis of noisy firing or other activation patterns that bore no meaningful representations according to what the model had learned about initially. Instead, when we decoded the hidden states of the model we found that the represented content after adaptation corresponded to the kind of images seen in training, whereas prior to adaptation, decoded images matched the empty input.

To quantify this, we measured hallucination quality (as defined in Section 4.2.2) over the course of homeostatic adaptation. In Figures 4.6(c, d) and 4.7(c, d), each dot represents the quality of the image decoded from the topmost hidden states at the end of the 50 sampling cycles in a trial. It becomes apparent that hallucinations started to emerge only after an initial period of silence, even as excitability was already adapting. This is consistent with cases reported in CBS where loss of vision was abrupt (Menon et al., 2003). The reported duration of this latent period, ranging from hours to days, in turn matches well the time scale over which homeostatic adaptation takes place (Turrigiano & Nelson, 2000).

In terms of quality, high-quality hallucinations were found soon after the point when hallucinations emerged (see Figure 4.8 for example decoded hallucinations[9]).

[9] As can be observed in the figure, the MNIST hallucinations of lower quality often looked like less well-defined digits or mixtures of different digit classes. Human judgement of quality and class could often deviate from the classifier's result in such cases. Similarly, what looked like a well-defined digit to us could still have features that were not present in the training set (for example, its precise positioning in the image), thus leading to a low score assigned the classifier.

Shapes data set



(a) Activity vs. iterations (200 or 700 in total).

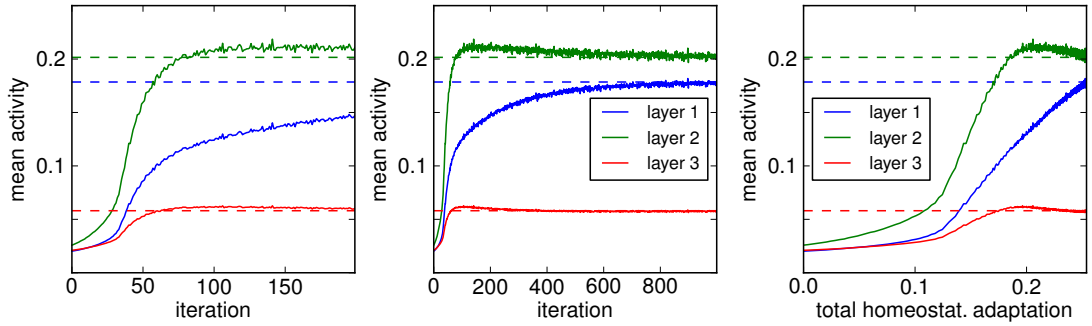
(b) Activity vs. bias adaptation.

(c) Hallucination quality (vs. iterations).

(d) Halluc. quality (vs. bias adapt.)

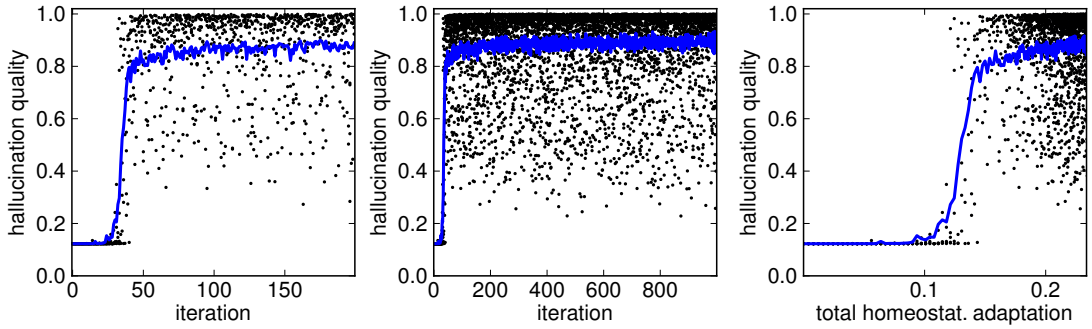
Figure 4.6: Emergence of hallucinations in a model trained on the shapes data set. *(a)+(b)*: with empty images as input, activity levels dropped in all three hidden layers and then recovered over the course of homeostatic adaptation (original levels as dashed lines; see Figure 4.4 for explanation of x-axes). *(c)+(d)*: quality of hallucinations (i.e. how well decoded internal representations matched the learned images, Section 4.2.2). Each dot represents the decoded internal state after the 50 sampling cycles constituting a trial (5 out of 100 trials per iteration are plotted). Blue curve denotes mean quality over 100 trials in that iteration. After an initial period of silence, hallucinations emerged abruptly, quickly rising in quality. The emergence of hallucinatory representations coincided with a more rapid recovery of activity levels.

MNIST data set



(a) Activity vs. iterations (200 or 1000 in total).

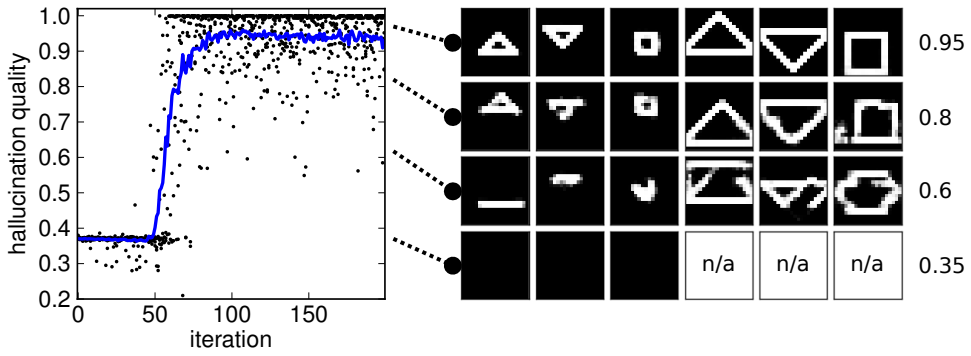
(b) Activity vs. bias adaptation.



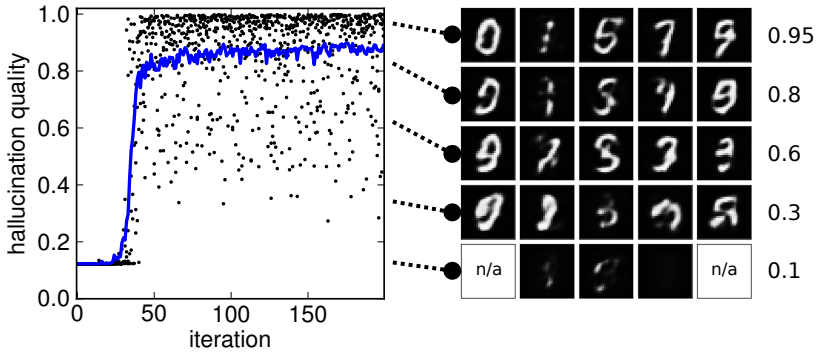
(c) Hallucination quality (vs. iterations).

(d) Halluc. quality (vs. bias adapt.)

Figure 4.7: Emergence of hallucinations in a model trained on the MNIST data set. *(a)+(b)*: activity changes. *(c)+(d)*: quality of hallucinations. Results are qualitatively similar to the model trained on shapes data (see Figure 4.6 and caption text for details). One difference was a slight initial overshoot of activity levels in the higher hidden layers (layer 2 and 3). Activities there then gradually dropped as the first hidden layer approached its original activity level from below.



(a) Shapes hallucinations.



(b) MNIST hallucinations.

Figure 4.8: Example decoded hallucinations (right-hand side), with corresponding scatter plots for reference (left-hand side, from Figures 4.6c and 4.7c). *(a)*: for the model trained on shapes, displayed are examples from the six shape categories (columns, as categorised by matching to the shape templates), for four different qualities (rows, with quality values listed on the right-hand side; images were of that quality or within +0.05 thereof). For entries marked 'n/a' there was no hallucination of that type and quality (note that the categories are not really meaningful for lowest quality images anyway). *(b)*: similar to (a), but for the model trained on MNIST. Examples shown were classified as belonging to digit categories 0, 1, 5, 7, and 9 (columns), for five different qualities (rows, annotation as in (a)). Note that the classifier result did not always agree with human judgement.

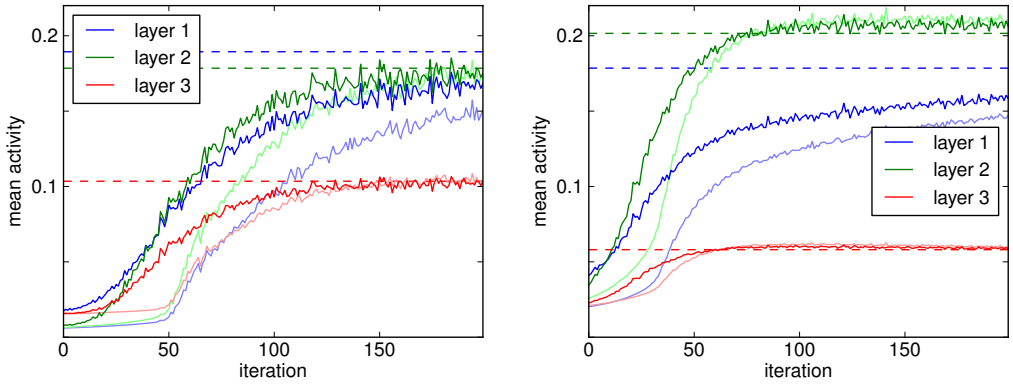
That point also marked a profound increase in the rate of activity changes. This shows that the emergence of stable internal representations is not just a epiphenomenon of underlying activity changes, but rather itself plays a key role in the system recovering normal activity levels.

Throughout the course of adaptation, we found there to be a mix of hallucinations of various qualities. Lower quality images could correspond to temporary states as the model transitioned from one relatively stable state to another. Note that within any one trial, the model never converges to a fixed internal state, as it keeps stochastically sampling from the posterior. We did observe a tendency to stay within one category of object (e.g. a specific class of digit) towards the end of a trial, but this is simply a general property of such models not specific to the hallucinations (namely, they do not mix well, to be addressed in Chapter 5). Similarly, hallucinations could come from various object categories (among the digit or shape classes) for an individual instance of the model. This matches reports from CBS patients, which indicate there can be a variety of hallucinatory content that varies from episode to episode for an individual subject. It is thus important that the model could produce varied representations rather than just a few degenerate states.

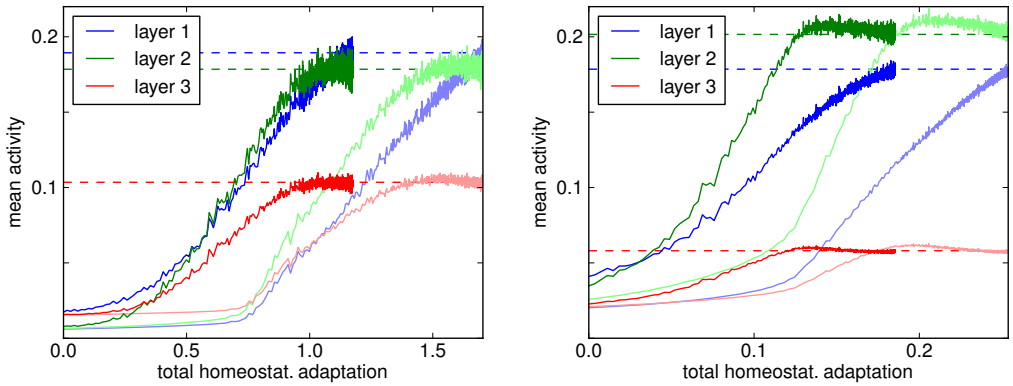
4.3.3 Sensory deprivation due to noise or impoverished input

The emergence of hallucinations in the model does not require complete lack of input. We obtained similar results when performing the homeostasis experiment with images containing, for example, some noise (10% white pixels on black background randomly sampled for each image). In that case, fewer iterations and less homeostatic adaptation were needed to trigger hallucinations (Figure 4.9). Hence, the nature of visual impairment can have an impact on when or whether hallucinations are occurring. This could also offer one possible explanation for why there might be a tendency for hallucinations in CBS to cease once vision is lost completely (Menon et al., 2003). If one assumes that there are limits to how much neurons can adapt their excitability, then some remaining input, even if it is just essentially noise, might be necessary to drive cortical neurons sufficiently. On the other hand, an alternative explanation for a cessation of hallucinations might be long-term cortical reorganisation or learning (see discussion, Section 4.4).

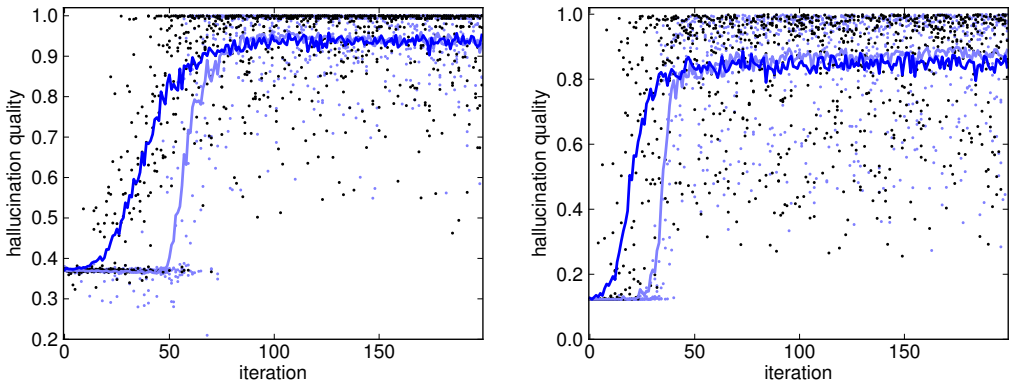
Still, one potential problem with our implementation of sensory degradation so far, be it with empty input or noise, could be that it corresponds to a rather extensive damage to the visual system. Perhaps one would be inclined to interpret such input degradation



(a) Activities (vs. early iter.) on blank or noise images, for shapes (left) or MNIST model (right).



(b) Activities (vs. bias adapt.) on blank or noise images, for shapes (left) or MNIST model (right).



(c) Hallucination quality on blank or noise images, for shapes (left) or MNIST model (right).

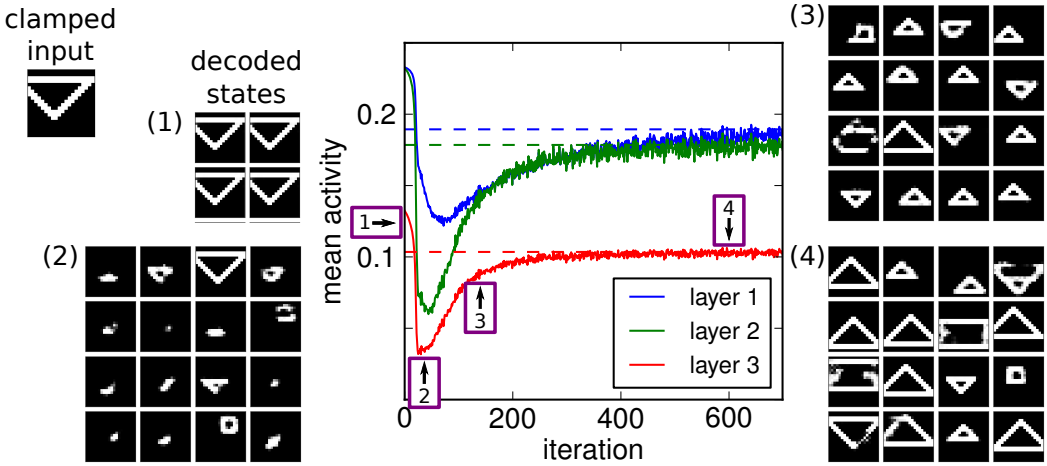
Figure 4.9: Comparison of results of homeostatic adaptation for blank input images (dark curves) or noise images containing 10% white pixels on black background (light curves). Left-hand figures are for the model trained on shapes, right-hand figures for the one trained on MNIST. With noise input, hallucinations emerged after fewer iterations and with less adaptation of the biases. Moreover, comparing the differences in activity between first and second layer across conditions, it appears that the recovery of the first was less delayed relative to the second when the former was receiving noise input.

as a model of complete blindness rather than a more graded visual impairment (or one that is more spatially restricted, Section 4.3.4), where in the latter case there might be some structure in the sensory data left. Moreover, in all experiments simulated so far, the emergence of hallucinations occurred due to homeostatic adaptation that compensated for a rather massive drop in activation levels caused by the lack of input. However, if the introduced homeostatic mechanism is truly effective at stabilising the *distribution* of learned internal representations, one could expect that the system could be prone to hallucinate under much more general conditions than just lack of input: as long as the ongoing input does not evoke a wide *variety* of learned percepts, those groups of neurons that participate in representing the lacking percepts might compensate by increasing their excitability, possibly causing corresponding hallucinations (where the definition of hallucination again would require that the internal representation deviates profoundly from what is in the external input).

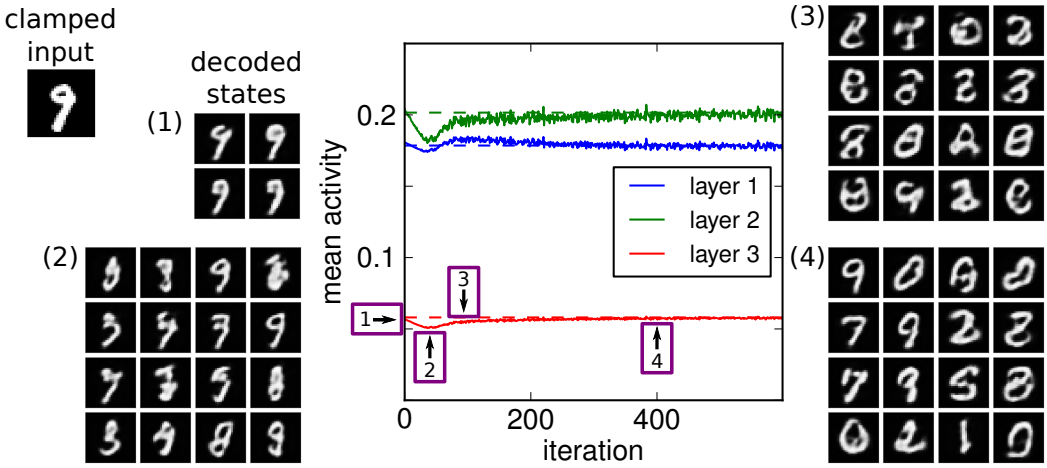
To address these issues, we aimed to test whether hallucinations were exclusively a consequence of compensation for overall lack of input and resulting activity decreases, or whether they could still emerge with structured input that was however highly impoverished in its variety. To this end, we simulated the homeostatic adaptation for the shapes and MNIST models, with the visible layer clamped to only a *single* fixed image from the respective data sets over the course of the whole experiment. To clarify, as before, this models slow neuronal changes over the course of perhaps days or longer, rather than fast neuronal adaptation during ongoing perception, with neuronal parameters being fixed during trials and only updated gradually between them.

Results are displayed in Figure 4.10, depicting activity changes over the three hidden layers and examples of decoded internal representations at various stages. We found that hallucinations did indeed develop: initially, the decoded internal states faithfully represented the image in the sensory input. However, as the neurons adapted over time to compensate for the impoverished input, the internal representations entailed objects not actually in the image, effectively decoupling perception from sensory input.

This result clarifies that the action of the homeostatic mechanism can be much more specific than just recovering overall activity levels. Indeed, for the fixed input images used, initial global activity levels when doing inference were actually at or above average, for the MNIST model and shapes model, respectively (as is shown in the figures). The homeostatic adaptation however acts locally for each neuron. With fixed input, one sub-population of neurons, whose activation distributedly codes for that input, will be highly active, while other groups of neurons are less active than average.



(a) Shapes hallucinations with impoverished input (fixed image).



(b) MNIST hallucinations with impoverished input (fixed image).

Figure 4.10: Modelling sensory deprivation due to impoverished input variety, we clamped the input layers of the models to a single image from the respective data sets throughout the course of homeostatic adaptation. Plotted are resulting activity changes and example decoded internal states, for the shapes *(a)* and MNIST *(b)* models. Initially, decoded images (1) corresponded to the input. As neurons adapted and the internal percepts deviated from the true input, global activities dropped (2), then recovered driven by hallucinatory percepts (3 and 4). For particularly the MNIST model, we also observed a more gradual improvement in hallucination quality (compare 3 to 4).

Adaptation of neuronal excitability can then continuously shift the balance, even if activity averages across a layer remain similar.[10]

As can be observed in the figures, there was an initial drop of global activity levels, especially for the shapes model. Based on the decoded representations at that point, we suggest that this results primarily from the neuronal population that represents the initial, veridical percept decreasing excitability. Then, as other neurons increase their respective excitabilities, alternative, hallucinatory internal representations take over, leading to a stabilisation of global activity levels.

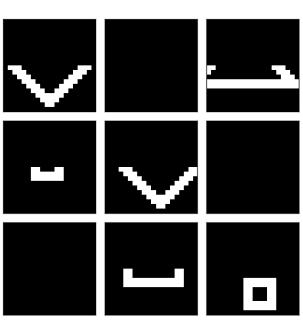
The degree of decoupling of the internal percepts from the sensory input was striking. It appeared to be surprisingly robust, overcoming not just a lack of input but even contradictory input. In the case of the shapes in particular, the hallucinated objects do not even necessarily share parts with the true input. It should be recalled that the homeostatic mechanism merely adapts the local biases, and thus does not at all change the connection strengths between units or layers. Indeed, we could show that the flow of information from sensory input to the higher layers was not completely prohibited in the model after homeostatic adaptation. Running a model that currently displayed hallucinatory representations, we modestly increased the impact of feedforward processing, using a mechanism meant to model the action of acetylcholine (Section 4.3.6). The internal representation then reliably realigned to the actual input image.

4.3.4 Localised and miniature hallucinations from localised impairment

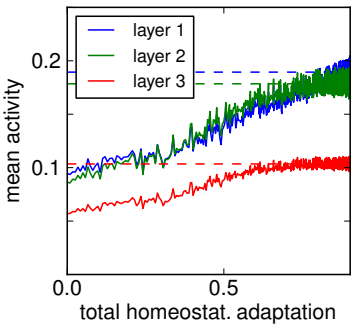
Visual impairment leading to CBS can also be constrained to specific parts of the visual field. Although reports are conflicting (Menon et al., 2003), for some patients at least hallucinations tend to be localised to these regions. We tested whether we could reproduce this finding using the model trained on the shapes data set, in which the objects are distributed across various image positions. We simulated a more localised impairment by repeating the homeostasis experiment while blanking only half of the images (for example the top half, Figure 4.11a). As before, the neurons' activities dropped initially and then recovered during adaptation.

In the original homeostasis experiment, where visual impairment involved the whole visible layer, hallucinated objects were distributed across the whole visual field (Figure 4.11d). However, when the model where only half of the images had been blanked

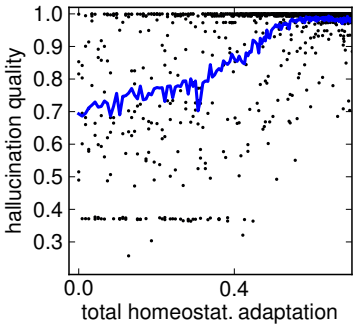
[10] We explore a related functional role of neuronal adaptation on shorter time scales in Chapter 5.



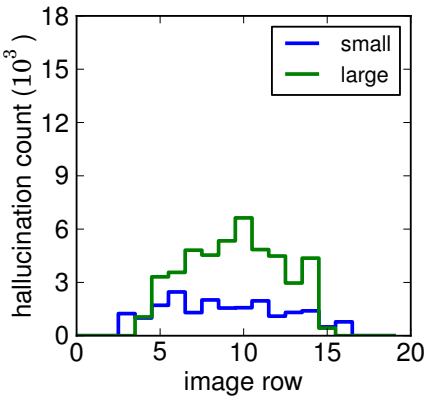
(a) Shapes, top half impaired.



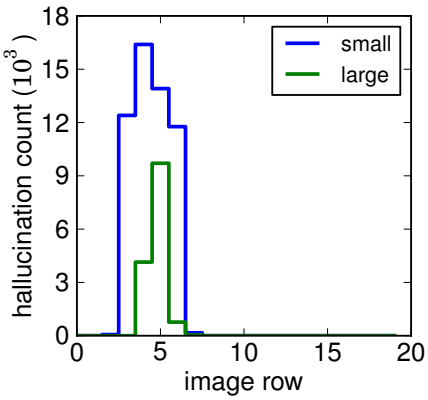
(b) Local. impairment, activities.



(c) Local. impairment, halluc.



(d) halluc. localisation, fully impaired.



(e) halluc. localisation, top half impaired.

Figure 4.11: Emergence of hallucinations in a shapes model where visual impairment was restricted to the top half of the input images. *(a)*: example images. *(b)*: recovery of activities during homeostatic adaptation. *(c)*: hallucination qualities during adaptation. Note that many of the corresponding decoded internal representations were not actually hallucinations, but rather matched shapes that were in the unimpaired half of the visual input. In particular, in the early phase of adaptation there are two clusters at low and high quality values. These correspond to void internal representations or veridical ones when shapes happened to lie completely in the impaired or healthy halves, respectively. The former then were gradually replaced with emerging hallucinations. *(d)*: distribution of hallucinated small and large shape categories across the image in model with fully impaired input. Only hallucinations with quality greater than 0.85 were counted here. *(e)*: as (d), but for the model that underwent adaptation with only the top half damaged (displayed data then taken with fully blank images as input as to not be influenced by actual objects in the healthy region). Now, hallucinations were localised to the impaired region and favoured smaller shapes, which would ‘fit’ within that region.

was tested (on blank images[11]), hallucinated objects were restricted to the image region that had been lesioned (Figure 4.11e). Excitability changes due to homeostatic adaption are thus specific enough in the network to have topographic properties.

Another occasional phenomenon in CBS is that hallucinated objects appear to be "Lilliputian" or miniaturised. It has been suggested that this can be explained as resulting from a mismatch of hallucinated content and context, where hallucinations appear against real visual background that happens to be too close in relation to the size of the hallucinated objects (ffytche & Howard, 1999). On the basis of our simulation results, we tentatively make another prediction: if there is a propensity for hallucinatory content to consist of meaningful wholes, such as full objects or faces, then in patients where hallucinations are restricted to impaired regions of the visual field there should be a correlation between object size and the spatial extend of visual impairment. To see this in our model, consider that in our shapes data set, objects could come either in small or large versions. For models with full loss of vision, hallucinations where biased towards the larger objects (Figure 4.11d).[12] On the contrary, in models with lesions restricted to the top half of the visual field, hallucinated objects were not only localised to the impaired region as reported above, but the frequency ratio was also reversed: smaller objects were much more common, and larger objects were less frequent and narrowly centred relative to the impaired region (Figure 4.11e). Moreover, we found that, without a single exception, *all* hallucinations of larger shapes happened to be of the 'downwards-triangle' category–the only large category where most of the object could fit into the lesioned region.

Thus, the process that generates hallucinations due to homeostatic adaptation can specifically evoke only certain types of content as determined by the nature of the visual impairment. Here, it is those objects that happen to fit within the boundaries of the lesion in the visual field.

4.3.5 The locus of hallucinations: cortical lesions vs. suppression

We then turned our attention to the question of the roles of different areas in the cortical hierarchy. As described in the introduction, the complex content of hallucinations

[11] Blank images were used for testing rather than the images with half the content removed in order not to conflate hallucinated content with content in the intact half.

[12] Possibly, this is because larger shapes evoked higher overall activity in the model and in turn were more suitable for activity restoration. Note for example in Figure 4.10a the transition from smaller to larger shapes as activity increases from point 3 to 4 (examples plotted there are for the model with impoverished input).

in CBS suggests the involvement of visual association cortex and other higher visual regions, and evidence implies that intact association cortex is both necessary and sufficient to develop complex hallucinations. For example, cortical lesions in early visual areas can bring about the visual impairment that causes complex hallucinations, but lesions that involve visual association cortex appear to prohibit them.

Interestingly however, a study by Merabet et al. (2003) suggests that lower areas, when at least partially intact, can still contribute to hallucinatory activity in an essential fashion. The authors examined a patient suffering from CBS due to visual impairment caused by lesions in early visual areas. Maybe contrary to expectation, applying Transcranial Magnetic Stimulation (TMS) to early areas in a way thought to cause cortical suppression lead to a temporary cessation of the hallucinations. The authors argue that their finding goes contrary to the 'release' theory of complex hallucinations, according to which the lack of input to higher areas from lower areas somehow disinhibits or releases perceptual representations in the former. Under this theory, the further suppression of the already damaged early areas in the patient should only have exaggerated the hallucinations.

We elucidate on these issues surrounding the role of areas in the cortical hierarchy using the DBM model. Of course, the hierarchy in the DBM, consisting of several hidden 'layers', is highly simplistic when compared to the specialisation and complex computations in cortical areas (as was discussed in Chapter 3). Nevertheless, it turns out that the aspect our model does capture, namely having several subsequent processing stages differentiated at least by increasing receptive field sizes, is adequate to explain the phenomena at hand.

To begin with, we found that DBMs trained without the topmost hidden layer failed to learn generative models of the data in the first place, and thus were inevitably incapable of producing corresponding hallucinations. This mirrors visual association cortex being necessary for complex hallucinations, and can be explained in the model with lower layers being incapable of learning the full structure of objects in the images, due to their limited receptive field sizes.

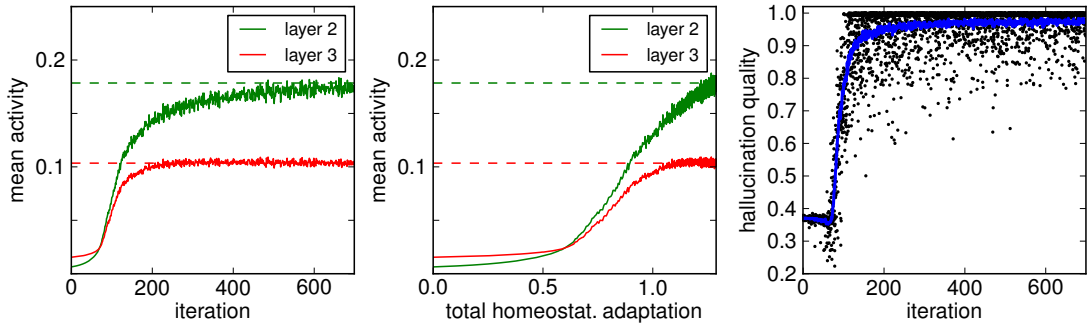
What about intact higher areas being sufficient for the emergence of hallucinations, while lower ones are not necessary? To model lesions to early visual areas, we repeated the homeostasis experiment, only this time we did not blank the input but rather 'lesioned' the first hidden layer, i.e. we clamped units in the latter rather than the units in the visible layer to zero (thus, with the first processing stage blocked, the actual content in the visible units was rendered irrelevant). As before, hallucinations did emerge over

the course of homeostatic adaptation (Figure 4.12). Hence, remaining layers in the model are sufficient in principle as long as they form a network that can synthesise the relevant information about visual objects.

Finally, we modelled the suppression of early visual areas with TMS in a CBS patient as described by Merabet et al. (2003). Unlike in the last experiment, where early areas were permanently incapacitated and higher areas adapted over time, the TMS experiment corresponded to a temporary suppression in a system that had already developed hallucinations (presumably due to prior adaptation to visual impairment). Our setup thus used a model that had undergone homeostatic adaptation in response to blank visual input but with all hidden layers intact,[13] as in Section 4.3.2, leading to hallucinatory activity. We found that when we then temporally clamped the first hidden layer to zeros, modelling suppression with TMS, hallucinations ceased. Thus, even though this 'early area' is neither sufficient nor necessary for the model to develop hallucinations in the long run (as shown earlier in this section), it can be essential for *ongoing* hallucinations if it was in the first place part of the system when it underwent homeostatic adaptation.

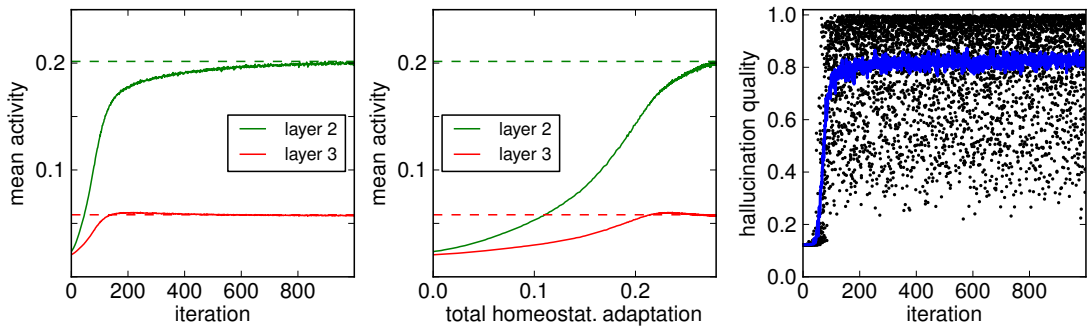
One possible interpretation of the relevance of lower areas could be that they provide higher areas with unspecific input, in the context of which the adaptation takes place. However, we suggest that the role of lower areas could be more subtle thanks to recurrent interactions with higher ones. As can be seen in the example in Figure 4.5, the representations assumed in lower layers during hallucinations are somewhat specific to the hallucinated object, even though those layers by themselves are incapable of synthesising it. Thus, this necessarily is a result of feedback from higher areas. It seems plausible that the lower areas could also contribute by stabilising the overall perceptual state assumed across the hierarchy. Then, any significant interference with representations in lower areas, not just suppression of activity, might impede hallucinations. Indeed, in the study of Merabet et al. (2003), even a TMS protocol not thought to cause suppression, when applied to primary visual cortex of the patient, resulted in a disruption of hallucinatory content. In the model, this could be tested by trying out different forms of manipulations other than suppression. At the time of writing we have not explored this further.

[13] We are assuming here that the lesions to early cortical areas in the CBS patient can be modelled to be upstream of the areas (or intact parts thereof) that were suppressed by TMS, even though all of them still counted as early in the visual system.



(a) Lesioned shapes model activities.

(b) Lesioned shapes model halluc.



(c) Lesioned MNIST model activities.

(d) Lesioned MNIST model halluc.

Figure 4.12: Emergence of hallucinations for the models that had their first hidden layer 'lesioned' (clamped to zero) rather than the visible input layer, modelling damage to early cortical areas rather than prior in the visual pathway. Results are overall analogous to the latter case.

4.3.6 A novel model of acetylcholine and its role in CBS

Finally, one relatively common feature among CBS patients is that hallucinatory episodes are more likely to occur in states of drowsiness or low arousal. This suggests a role of cholinergic systems, which in turn are implicated in complex hallucinations in a variety of situations outside of CBS, whether drug induced or disease related (Perry & Perry, 1995; Manford & Andermann, 1998). Indeed, in the (non-computational) model of complex hallucinations by Collerton et al. (2005), acetylcholine (ACh) dysfunction is attributed a major importance. At the same time, there is no evidence that an actual ACh dysfunction exists in CBS. Rather, in CBS the correlation with state of arousal might be effected by an interplay of hallucinations with physiologically normal fluctuations of ACh.

Making the connection between a lack of ACh and hallucinations is natural as there is experimental evidence that ACh acts specifically to emphasise sensory input over internally generated input, mediating "the switching of the cortical processing mode from an intracortical to an input-processing mode" (Sarter et al., 2005). In the computational model of Yu & Dayan (2002), ACh is modelled in a Bayesian framework to modulate the interaction between bottom-up processing carrying sensory information and top-down processing conveying prior expectations. The authors noted the relation to hallucinations, but to our knowledge, there is no computational model exploring it concretely.

Here, we explore an extended interpretation of the action of ACh as mediating the balance between external and intracortical input: in the hierarchy of cortical areas, ACh could affect the balance in the integration of feedforward and feedback information at each stage of the hierarchy. At an intermediate stage, feedforward information from lower areas indirectly carries sensory input, and feedback information is more internally generated, keeping with the idea of a ACh mediated switch between external and internal inputs. However, both feedforward and feedback input would in this case be intracortical (perhaps with additional effects on any direct thalamic inputs).

We thus model the effect of ACh in the following way. In the DBM model, each (intermediate) hidden layer receives input from a layer below, conveying indirectly sensory information, and from a layer above that has learned to generate or predict the former layer's activity. ACh is to set the balance in between feedforward and feedback flow of information. We introduce a balance factor $\alpha \in [0, 1]$, so that an intermediate

layer $\mathbf{x}^{(k)}$ is sampled as

$$P(x_i^{(k)} = 1|\mathbf{x}^{(k-1)}, \mathbf{x}^{(k+1)}) = \sigma(\sum_j 2\alpha w_{ji}^{(k-1)} x_j^{(k-1)} + \sum_j 2(1-\alpha) w_{ij}^{(k)} x_j^{(k+1)}), \quad (4.2)$$

given states $\mathbf{x}^{(\cdot)}$ and weights $\mathbf{W}^{(\cdot)}$ above and below (biases omitted for brevity). Hence, $\alpha > 0.5$ corresponds to increased feedforward flow of information, assumed to model increased ACh levels, and $\alpha = 0.5$ recovers the normal sampling mode for normal levels. We note that this mechanism is a heuristic in that it treats the DBM as a neural network more than a well-defined probabilistic model. In particular, for $\alpha \neq 0.5$, the effective connections between layers are no longer symmetrical and thus the model no longer constitutes a Boltzmann machine.[14] The resulting model of the action of ACh differs in several regards from that of Yu & Dayan (2002), other than applying to several stages in a hierarchy. We will briefly comment on these differences in the discussion.

ACh and contour completion

Before we turn to the role of ACh in CBS, we first briefly demonstrate its effect on the balance between feedforward and feedback in the model under normal sensory input. One example where it has been suggested that feedback could play a role is contour completion, for instance as framed in the hierarchical Bayesian inference account of Lee & Mumford (2003). Given an incomplete stimulus, higher areas might fill in missing information and subsequently convey it to lower areas, possibly leading to the perception of illusory contours.

We explored this phenomenon and a possible interaction with cortical ACh levels by testing the models on modified images where parts of the objects had been blanked out (Figure 4.13). Shown are examples of the decoded representations inferred by the model, for all three hidden layers and three different levels of ACh in both intermediate hidden layers, each for two different input images. We found that completion did indeed take place, especially in higher layers. Increased ACh levels, modelled with $\alpha = 0.7$, resulted in an emphasis on bottom-up processing, leading to less completion, in particular in lower layers that now received less top-down feedback. Decreased ACh levels on the other hand had the opposite effect.[15]

[14] In a sense, the factor α interpolates between inference in a DBM and approximate inference in a deep belief net (Section 2.2.2) defined with the same parameters.

[15] It should be noted that the generative nature of the DBM allows for much more extensive completion of image information if the visible units themselves are sampled for pixels where filling in should take place. For example, Eslami et al. (2012) showed that a small region of clamped visible units can successfully constrain generation in the rest of the image. In our example however, the whole visible layer

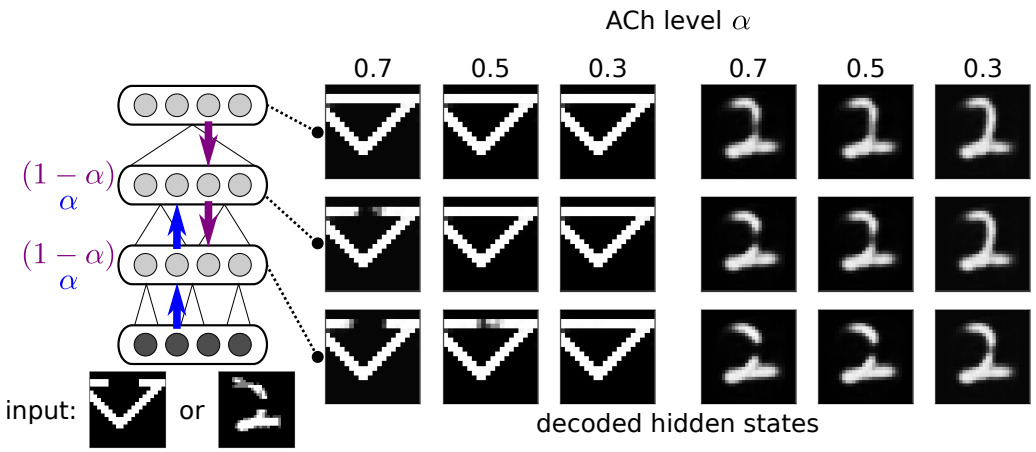


Figure 4.13: Contour completion and interaction with ACh levels α in the model. Incomplete images (lower left) were given to either shapes or MNIST models as input. Displayed are decoded hidden representations for the three hidden layers (rows), for three different levels of ACh (columns). Mean-field inference (i.e. propagating activities instead of samples, Section 2.2.2) was used here to reduce sample variability/noise. Filling in of missing contours occurs more in higher than lower layers. ACh shifts the balance towards bottom-up processing, leading to less filling in with increased levels.

ACh and CBS

We modelled the effect of drowsiness or low arousal on hallucinations in CBS as follows. We assumed that drowsiness is accompanied by a decrease in ACh, modelled as $\alpha = 0.3$. This value was chosen to obtain a clear effect while still allowing for both feedforward and feedback processing to play a role during inference. The precise value however did not matter much. As states of drowsiness are intermittent with periods of normal or increased vigilance (there is no known pathology of these aspects in CBS per se), we assumed that on average, ACh levels are still balanced. Hence, the homeostasis experiment was conducted such that at each iteration, half of the trials were performed with low α, and the remainder with increased ACh levels at $\alpha = 0.7$, yielding a normal value of 0.5 on average.

For both shapes and MNIST models, results are displayed in Figure 4.14, both for trials with $\alpha = 0.3$ and $\alpha = 0.7$ (dark and light curves, respectively). We found that with decreased levels of ACh, less homeostatic adaptation of excitability was necessary

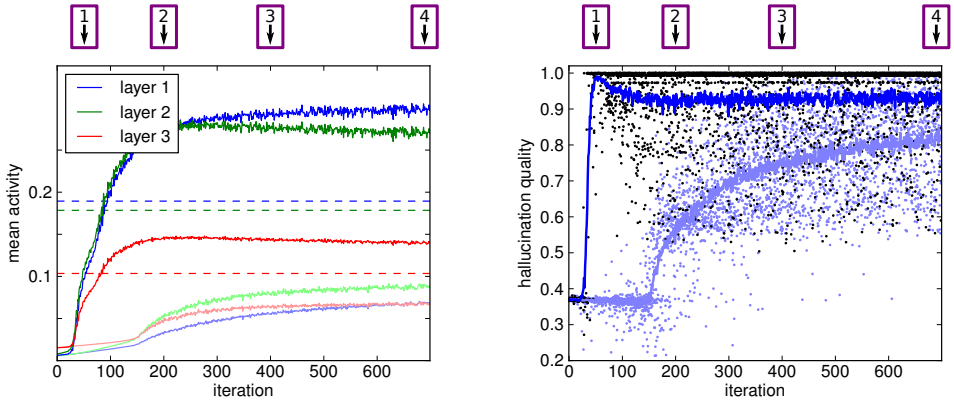
always remains clamped, and filling in only happens in the subsequent hidden layers. This is because in our model, the visible layer is meant as an early stage of processing where input is still faithfully represented in a purely bottom-up fashion. We also interpret contour completion as a more graded effect that happens throughout the hidden representations, rather than completely surmounting the input itself.

to elicit hallucinations (adaptation values not shown in figures for brevity, but they correlate with iteration number as in all experiments before). For example, for some intermediate level of adaptation, hallucinations only occurred with decreased but not with increased levels of ACh. This would thus correspond to a situation where hallucinations would only be triggered during drowsiness. Later on, hallucinations started to appear for $\alpha = 0.7$ as well, but they generally were weaker and less formed for most of the simulation duration.

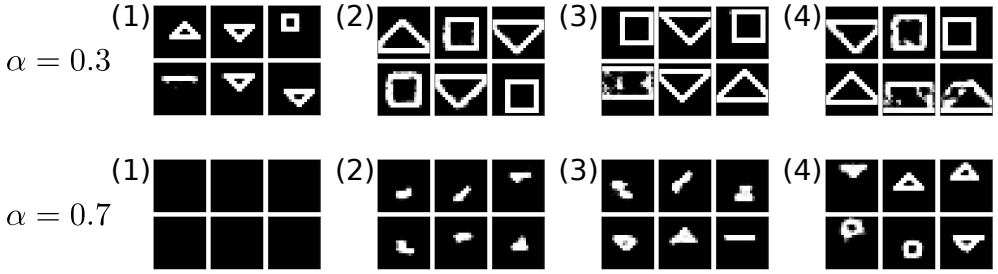
In terms of neuronal activity levels, increased occurrence of hallucinations for low ACh levels corresponded to increased activity in the hidden layers. In particular, throughout later parts of the simulation, average activity levels for each hidden layer were generally twice as high during trials with $\alpha = 0.3$ compared to those with $\alpha = 0.7$, meaning that on average original activity levels were restored. Note that, as was described in the methods (Section 4.2.2), each iteration consisted of 100 trials over which current activity was averaged for each neuron and then compared to the original values. This result demonstrates that intermittent hallucinatory episodes alternating with silent periods, as is often the case in CBS, could restore mean activity levels, as long as the timescales over which neurons measure their average activity are long enough to encompass both. A possible prediction from our findings could be that cortical activation during hallucinatory episodes should actually be higher than what they had been during healthy perception.

In terms of hallucination quality, there was actually a peak very early on for low ACh, coinciding with the point in time when activity levels within low ACh trials crossed approximately their original levels (examples at point (1) in the figures). Because the neurons however measured current activity over both low and high ACh trials, activity for the former had to increase further, leading to a decreased quality of hallucinations. This was especially true for the MNIST model, where unnaturally high activity resulted in over-expressed imagery that showed little variety (Figure 4.14c (2)). However, over the further course of adaptation in the MNIST model, activity levels for low ACh dropped again somewhat as the trials with higher ACh began to contribute activity, resulting in more distinct if still somewhat over-expressed hallucinatory images (Figure 4.14c (4)).

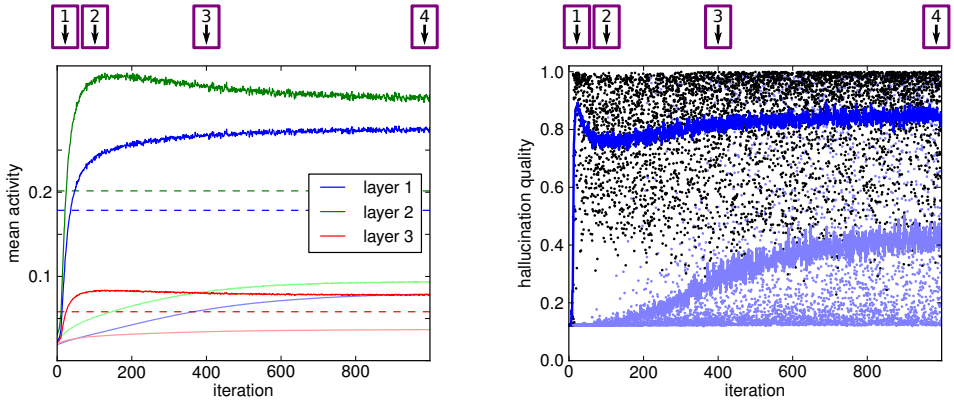
A related finding was a relationship between global activity levels and hallucinatory content (beyond just quality) in the shapes model. Corroborating what we observed earlier (Footnote 12 on page 86), internal representation of smaller shapes evoked less activation (averaged over a hidden layer) than that of larger shapes (compare halluci-



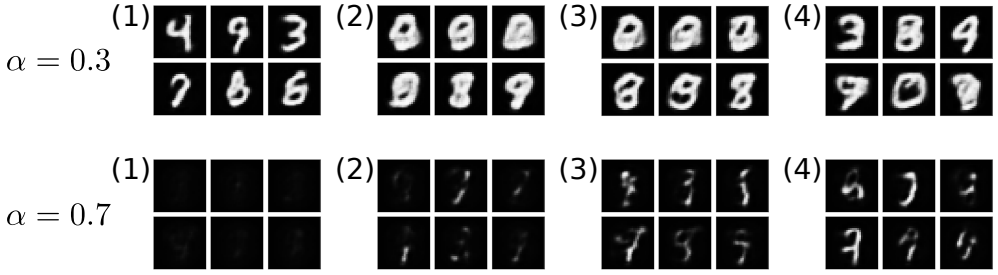
(a) Shapes model activities and halluc. qual. for ACh level $\alpha = 0.3$ (dark) and $\alpha = 0.7$ (light).



(b) Shapes model example hallucinations.



(c) MNIST model activities and halluc. qual. for ACh level $\alpha = 0.3$ (dark) and $\alpha = 0.7$ (light).



(d) MNIST model example hallucinations.

Figure 4.14: Caption see next page.

Figure 4.14: Emerging hallucinations over the course of homeostatic adaptation when each iteration consisted of both trials with low and high values of the ACh parameter, $\alpha = 0.3$ (dark curves) and $\alpha = 0.7$ (light curves), respectively. *(a)*: average activities and hallucination quality for the shapes model. *(b)*: example decoded hallucinations at time points indicated in (a). *(c+d)*: analogously for the MNIST model. For both models, lower ACh levels led to hallucinations earlier and with less homoestatic adaptation (only iterations shown here). In particular, there is an early phase in which hallucinations occurred only with $\alpha = 0.3$. Hallucinations that do emerge later on for $\alpha = 0.7$ remain weaker and less formed for most of the simulation. The difference in frequency of hallucinations also entails a corresponding difference in activity levels.

nation examples in Figure 4.14b with the activity levels at the corresponding points in Figure 4.14a). Because alternating ACh levels forced the system to restore average activity levels by relying on hallucinatory episodes of heightened activity intermittent with relatively hallucination-free periods of lower activity, well-formed hallucinations developed to be mostly shapes of the larger categories. The homeostatic need for increased activity levels during hallucitanory phases thus led to hallucinations of larger extent in the shapes model and over-expressed digits in the MNIST model. Possibly, such over-activation of cortical neurons might explain why hallucinations in CBS can be so vivid, for example involving "hyperintense, vivid, brilliant colours" (ffytche et al., 1998).

In summary, we found that a temporary change in the balance of feedforward and feedback flow of information can have a profound effect on the emergence of hallucinations, yielding a potential explanation for the role of drowsiness and ACh in CBS.

4.4 Discussion

We modelled the emergence of complex hallucinations in CBS as a result of homeostatic regulation of neuronal firing rate in response to degradation of visual input. Our computational model thus elucidates on similar suggestions in the literature (Burke, 2002; Plummer et al., 2007). The homeostasis mechanism is meant to underlie specifically CBS. Other pathologies involving complex hallucinations, such as schizophrenia or Lewy body dementia (Manford & Andermann, 1998), might have different causes. In particular, it might not be feasible to unify complex hallucinations in a single explana-

tory framework (cf. Collerton et al., 2005). What different conditions accompanied by complex hallucinations do have in common however is that they show that the brain can spontaneously synthesise rich representations of visual imagery, even in absence of or in contradiction to actual sensory data. Following notions of the brain implementing perception as analysis by synthesis, our study makes use of the DBM model that can learn to synthesise internal representations of images, in an unsupervised fashion, by virtue of being a generative model.

We reproduced a variety of qualitative aspects of CBS found in some patients, such as an initial latent period, a possible localisation of hallucinations to impaired parts of the visual field, and the effect of suppression of cortical activity. We suggest that interfering with cortical homeostatic mechanisms might prevent the emergence of hallucinations in CBS. We also predict a possible correlation between a tendency to experience miniature versions of objects and the degree to which the spatial extent of visual impairment is limited, as well as activity levels during hallucinatory episodes possibly being higher than what they had been during comparable, stimulus evoked normal perception. We introduced a novel model of the action of acetylcholine (ACh), suggesting that it could not only influence the balance between thalamic and intracortical inputs (Sarter et al., 2005), but also the balance between feedforward and feedback at various stages of the cortical hierarchy. In CBS in particular, a possible lack of ACh at cortical sites, e.g. during normal fluctuations entailed in changes of state of arousal, could be conducive to the emergence of hallucinations.

In the model, internal representations of learned objects were robustly recovered by the homeostatic adaptation in a variety of conditions, be it complete lack of input, noise input, or naturally structured but highly impoverished input consisting of fixed images. A key aspect of the model was that hallucinations did not consist only of stereotyped images, but rather a variety of percepts reflecting at least a part of the full distribution of objects learned initially. This variability was due to different groups of neurons participating in coding for different percepts, meaning that a local homeostatic restoration of activity levels for the population required activation of a variety of percepts over time.

That hallucinations emerged even when normal input images were used but kept fixed over the course of homeostasis, shows that it was not so much the total lack of sensory input or global drop in evoked activity that mattered, but rather the failure of the given input to evoke a wide *range* of learned percepts. Whether impoverished input can have such a powerful impact on perception in reality should be explored further. There is indeed evidence that sensory deprivation (in terms of general impoverishment, not just

complete lack of sensory input) can cause hallucinations in healthy individuals (Corlett et al., 2009; Menon et al., 2003), but there seems to have been little experimental work along that direction since the sixties (Mason & Brady, 2009).

To our knowledge, our work constitutes the first computational model that concretely explored aspects of CBS. Other neurological pathologies have been studied before with neural network models (Finkel, 2000; Aakerlund & Hemmingsen, 1998). Probably most closely related to our work, Ruppin et al. (1996) modelled the emergence of hallucinatory memory patterns in schizophrenia, using a Hopfield network (a line of work initiated by Hoffman & Dobscha, 1989). The underlying mechanism, homeostatic plasticity in response to input degradation, is quite similar, and they made some analogous observations, including a beneficial role for homeostatic regulation for stabilising neuronal representations. However, in their model the hallucinatory 'memories', supposedly residing in prefrontal cortex, are accounted for much more abstractly, consisting of random patterns. Moreover, the retrieved patterns in a Hopfield net correspond directly to the patterns provided as input. It is thus not obvious how to relate their network and the stored patterns to specifically visual processing, which is essential for studying CBS. Our model can be seen as a significant extension of their work in that direction. It involves hierarchical, topographic representations of images, learned in a generative model framework. In particular, the synthesised representations are interpreted to play an integral part in *perception* itself, not just in unspecified memory-like pattern recall. A generative model moreover relates to other approaches discussed in the context of hallucinations (Bayesian inference, predictive coding, adaptive resonance; Yu & Dayan, 2002; Corlett et al., 2009; Friston, 2005; Grossberg, 2000).

We emphasise the distinction between the roles that homeostatic adaptation and learning play in our model and possibly the cortex. Learning is to be seen as a lasting change of circuitry that captures aspects of the sensory input in the neuronal representations, improving the network's function according to some criterion. In the generative model, that criterion would be the ability to generate or predict the input itself, but it could also be the utility of the representations towards some other goal, such as discrimination of objects. Homeostatic adaptation on the other hand could serve to stabilise neuronal representations. While such stabilisation can in turn be important during learning itself (Turrigiano, 2008), we have shown in the model that it could offer a simple local mechanism to make representations more robust once they have been learned, for instance to counteract degradation in input quality (Ruppin et al., 1996)–thus effectively resisting changing aspects of the input, rather than capturing them via learning.

At the point in time where we simulate homeostatic stabilisation, learning might have concluded, having taken place in earlier stages of development, or it could still occur but over longer time scales. A decoupling of the time scales of homeostatic adaptation and learning could also explain why CBS can recede over time. Hallucinations might initially be caused by the short-term homeostatic regulation of neuronal activity, but long-term cortical reorganisation could lead to their cessation (Burke, 2002). In our framework, such reorganisation would correspond to learning to generate the *impaired* sensory input. Indeed, if we continue actual learning in the model as the input layer is clamped to empty or noise images, rather than just perform homeostatic adaptation, the model learns to generate and thus represent the empty input, losing the capability for hallucinations in the process.

4.4.1 Some open questions in CBS

One of the issues we have not addressed is what limits the incidence of complex hallucinations and CBS to about 11% to 15% of patients suffering from visual impairment (Menon et al., 2003). Our modelling results suggest however that a variety of parameters can influence whether and when hallucinations occur. In the model, the nature and degree of visual impairment as well the effect and variability of other interacting factors, such as ACh levels, determine how much homeostatic adaptation is necessary to push cortical activity into the hallucinating regime. Limits on how much cortical neurons can adapt their excitability therefore would restrict hallucinations to only certain cases, and there might be variability in such parameters of homeostasis across the population as well. Thus, that only some patients with visual impairment develop hallucinations could simply reflect the variance of the underlying relevant parameters. Similar reasoning might explain the diversity of symptoms among CBS patients.

Differences in hallucinatory content, e.g. whether it does or does not involve movement, faces, strong colours, etc., likely relate to the specialisation of different cortical areas (ffytche et al., 1998; Santhouse et al., 2000), and potentially to their selective sensory deprivation (such as more extensive impairment of colour vision possibly predisposing patients with senile macular degeneration to experience coloured hallucinations, Santhouse et al., 2000). A specialisation of different areas to different aspects of the sensory data was not a feature of our model. However, it seems reasonable to extrapolate from our results to a model extended in that regard. In our simulations, restricting sensory input by either removing only parts of the images or by just fixing

input to a single image led to hallucinations that reflected the specific lack in the input (namely hallucinations in the deprived part of the visual field, or of object types not present in the fixed input image, respectively). If different parts of the model were to distinctly represent properties of visual input in analogy to for example cortical areas V4 for colour and MT for motion, we would expect a specific deprivation of that input property to lead to corresponding hallucinatory representations.

An open question in CBS is also in how far hallucinated content reflects visual memories of some sort (Menon et al., 2003), although the elaborate and occasionally bizarre nature of the images might speak against this (see Teunisse et al., 1996; Plummer et al., 2007, for examples). In this context it is relevant that the DBM has been shown to be capable of synthesising images that generalise beyond what it has been trained on (Eslami et al., 2012). Moreover, in light of the bizarre or unusual hallucinatory imagery in CBS, some hallucinations with low quality in our simulations (as measured relative to training images) could possibly be interpreted as such unnatural imagery (see e.g. Figure 4.10b (3.); Ruppin et al., 1996, made a similar observation in their model).

4.4.2 Challenges for a computational model of CBS

The key for a model of CBS is to account for the ability of the brain to synthesise rich internal representations of images even without visual input, representations that possibly generalise over earlier experienced inputs (as argued above). This does not *necessarily* imply that the brain implements a generative model, in the sense captured by the DBM. However, the strength of such generative frameworks is that they account for these aspects naturally, at least in principle.

For comparison, a perceptual Bayesian model defined over a single low-dimensional variable can be sufficient to account for perceptual *illusions* concerning a property of an object (as does a prior for slow speeds, Weiss et al., 2002), but it is far-off from actually generating a full visual representation of the object itself. Similarly, the necessity for synthesis without input implies that a model computing a rich *code* of a given image is on its own not sufficient either. For example, the predictive coding model of Rao & Ballard (1999) and the sparse coding model of Olshausen & Field (1997) are both formulated as generative models that learn representations from images. Given an input image, they can infer a code that is rich enough in information to reconstruct the former. However, neither model can, when run purely generatively, synthesise structured images or anything akin to objects (although Rao & Ballard, 1997, demonstrate that memorised

images can be recalled). In particular, sparse coding trained on images tends to discover localised patches of edges as independent 'causes'. Thus, without an extension to higher level causes, a generated image will be a random superposition of such edges.

Similarly, neural networks like (deep) auto-encoders learn internal representations by reconstructing input. Using bottlenecks in the hidden layers, sparsity, input reconstruction from noise-corrupted input and other techniques (Bengio, 2009), they also learn about the underlying structure in images, enabling them to reconstruct from corrupted input, perform dimensionality reduction, or even learn transformations of the content (Hinton et al., 2011). However, there is no way of generating from these models in the absence of input (but see the recent work of Rifai et al., 2012). Hence, again such an approach might be used to model illusions, but not hallucinations.

Clearly, while our model, the DBM, is a generative model, its capability to generate 'images' still leaves much to be desired when it comes to matching the perceptual richness attributed to real images (although the DBM and closely related models have shown more potential in that regard than what is demonstrated here, see Salakhutdinov & Hinton, 2009; Ranzato et al., 2011; Courville et al., 2011). As model of cortical representations and processing, it also makes several simplifying abstractions, such as lumping together the highly differentiated feedforward and feedback connections in the cortex into simple symmetrical connections. This was discussed in detail in Chapter 3. Of particular interest are thus recent extensions that could enhance the generative performance of DBM-like approaches while at the same time having biological relevance as well, such as including lateral connections (Osindero & Hinton, 2008) or complex cell like pooling (Lee et al., 2009; Ranzato et al., 2010).

However, the results of this chapter demonstrate that the DBM does in principle capture several aspects important for our explanation of CBS. It is not meant as definitive model of generative processing in the brain, but rather serves as a simple idealised model system just complex enough to convey the points in question. Among the relevant aspects it captures is, first, the aforementioned capability to synthesise representations of input. Second, its hierarchical and topographic representations allowed us to model localised impairment and a role for acetylcholine. Third, the nature of the DBM as a neural network made it possible to model concrete cellular homeostatic mechanisms. Fourth, unlike for example the earlier Helmholtz machine model (Dayan et al., 1995), the DBM uses top-down interactions also during inference, not just learning, another requirement for modelling the role of hierarchical bottom-up and top-down processing for hallucinations.

There are not many concrete computational models of cortical processing that capture these aspects, but some related approaches do exist and will be considered in the discussion chapter (Section 7.1).

4.4.3 ACh and probabilistic inference

Our model of the action of ACh is closely related in spirit to that of Yu & Dayan (2002). In a sense we addressed some of the issues they identified with their own approach, namely only dealing with a localist representation of a low-dimensional variable, and only with a shallow hierarchy where the interaction of bottom-up and top-down is confined to a single stage. As they write, "it would be more biologically realistic to consider distributed representations at each of many levels in a hierarchy", which might be closer to what our model implements. However, the functional role of ACh was not the main focus of our work, and in some ways their model is significantly more sophisticated than ours.

Part of the issue is that our implementation is not necessarily a principled one within the DBM framework, and more of a heuristic. On the other hand, it could be argued that in some ways this is not too different from Yu & Dayan's approach in that regard, as they similarly characterise the ACh mechanism as implementing an approximation to exact inference. The latter crucially relies on only a single hypothesis being maintained at any point in time by the top-down part of the system, so that ACh controls the impact of that hypothesis on perceptual inference, which otherwise is driven by bottom-up sensory information. This is thus comparable to the action of ACh on the influence of higher layers on lower layers in our model.

However, where Yu & Dayan's model clearly goes beyond ours is that there the ACh level is itself controlled by the system dynamically during ongoing inference, whereas we merely manipulated it manually to explore its impact on emerging hallucinations. Whether such an internal control of the ACh parameter α could be implemented in the DBM framework is open. Again, the challenge might be not so much to come up with a corresponding heuristic mechanism, but to implement one that does not give up on the principled mathematical treatment of the DBM as a statistical model.

Another issue is in how far the role of ACh is necessarily to be interpreted in 'Bayesian' or probabilistic terms. Due to the approximations made in Yu & Dayan's model, the role of the high-level posterior for inference is reduced to an influence of a single current hypothesis and its associated uncertainty. The interpretation of the

effective impact of the top-down hypothesis as relating to its *uncertainty* seems justified as the latter is itself subject to ongoing probabilistic inference in their model. Because a mechanism for inferring this uncertainty is lacking in our model, we would be more cautious to necessarily frame the interaction of bottom-up and top-down as 'Bayesian' here. Similar applies to whether the overall issue of hallucinations should necessarily be framed in Bayesian terms. For our approach at least, the probabilistic nature of the DBM here only comes into play in so far as it allows for a means of formulating and deriving a generative model of sensory data.

4.4.4 The nature of hallucinatory experience

A subtle issue is how much information needs to be synthesised in the brain, and in what form, to generate the visual experience of hallucinations. Ultimately this line of enquiry leads to the deep question of the nature and the neural correlates of consciousness (Lamme, 2006), which we cannot hope to answer nor really address extensively here. We can however attempt to pose conditions for the generated neuronal representations necessary (though not necessarily sufficient) for evoking complex visual hallucinations: they somehow must entail the information content that is implied in the percept. For example, the experience of seeing, and presumably of hallucinating, a dog entails much more than just the information of the object in question being indeed a dog, i.e. some sort of category label. Rather, it involves perceiving the shape, contours, texture, colours, and so forth. Assuming the reports of CBS patients are not just confabulations, all the detailed information entailed in their hallucinatory percepts needs to be accessible to them from the underlying internal neuronal representations.

As an example of a model that would be clearly insufficient to synthesise the necessary information, consider a simple perceptual model consisting of a neural network classifier such as a perceptron, which has learnt to classify images of dogs against other images, using a single binary output 'neuron'. Internal activation of this unit alone cannot possibly be accompanied by the visual experience of seeing a dog, as the state of the output unit cannot be used to differentiate among the various possible instantiations of dogs (a dalmatian in a specific pose rather than a poodle in another, etc.). Nor can the single bit of information entailed in its state possibly convey all this detailed information. See Tononi (2008) of a discussion of such issues in the context of a theory of consciousness.

On the other hand, surely the generation of a hallucinatory experience does not

entail synthesising literal images in the brain, pixels and all,[16] but rather some form of representation thereof. In the context of hierarchical Bayesian models, a generative model is usually understood to involve a process that generates a lower level representation of the data from some variables that are by some measure more high-level, more abstract, etc. With regards to hallucinations, all we can say is that the relevant perceptual information is present and accessible to the subjects as to be available for report. The precise nature of the form of representation might be more difficult to discern. In particular, it is unclear whether a re-entrant top-down process is necessary to generate conscious experience (Lamme, 2006), or whether synthesis could be localised to high areas, thus not requiring a top-down generative component. Still, what makes generative models interesting is that they offer ways for how to learn rich internal representations from the data in the first place.

4.4.5 Conclusion

In this chapter, we showed how the DBM as a generative neural network can provide potential insights into how complex visual hallucinations emerge in CBS. By definition, such hallucinations rely on internally synthesised representations that deviate from actual sensory input. This notion will be also explored in the remaining two result chapters, but in a different light: the active synthesis of internal representations beyond what is determined by the senses can also play a functional role, either, to explore perceptual alternatives when input is ambiguous, evoking bistable perception (Chapter 5); or, to assume an internal state that specifically represents an object of interest, implementing object-based attention (Chapter 6). In the chapter on bistable perception, we will also examine neuronal adaptation again, but on much shorter time-scales than that of homeostatic regulation of firing rate, the latter possibly causing hallucinations in CBS.

[16] Although this could be possible in principle if there was more extensive feedback to the retina.

Chapter 5

Probabilistic sampling and neuronal adaptation in bistable perception

As we have seen in Chapter 4, hallucinations could exemplify the capability of the brain to synthesise internal representations of sensory input. They constitute an extreme case of a pathological decoupling of internal percept and external world. But what is the functional role of internal synthesis under normal conditions?

In a hierarchical generative framework, synthesising internal representations and employing them to make predictions about sensory input makes it possible to evaluate one's internal model and use this information to learn about structure in the data. The parameters of the model can thus be acquired in an unsupervised fashion. But a synthesis or generative component of a model is not always used for perceptual inference outside of learning, as the example of the Helmholtz machine (Dayan et al., 1995) shows, which uses an approximate feedforward recognition model during inference. On the other hand, as discussed in Chapters 1 and 4, analysis by synthesis accounts do attribute an important role to internal synthesis for perception itself. In particular, when sensory input is ambiguous, higher areas in the cortical hierarchy might communicate information about a high-level hypothesis to lower areas via feedback, in order to evaluate that hypothesis or to help resolving more ambiguous low-level hypotheses represented there (Mumford, 1994; Hinton et al., 1995; Carpenter & Grossberg, 1987; Lee & Mumford, 2003; Yuille & Kersten, 2006).

One possibly related and well-studied perceptual phenomenon is that of bistable perception[1] (Leopold & Logothetis, 1999; Tong et al., 2006; Sterzer et al., 2009). Here,

[1] The more general term would be *multi*stable perception, or perceptual multistability, for when there are more than two main interpretations, but all cases discussed here are of the binary type.

a subject's percept of an unchanging visual input switches over time. The input can be an inherently ambiguous image such as the Necker cube (Figure 5.1a) that allows for two different interpretations equally consistent with the sensory information, and the subject's percept alternates between these interpretations. Or, in binocular rivalry, conflicting images are provided to either eye, again eliciting a perceptual switching among the alternatives. Bistable perception might reflect how the brain deals with uncertainty posed by the ambiguous input, and is thus a test case for Bayesian or probabilistic accounts of perceptual processing (Bialek & DeWeese, 1995; Dayan, 1998; van Ee et al., 2003; Sundareswara & Schrater, 2008; Hohwy et al., 2008; Gershman et al., 2009a; Moreno-Bote et al., 2011; Gershman et al., 2012). Such approaches could offer normative alternatives to more classic explanations of perceptual bistability that model it mostly as an epiphenomenon of low-level neuronal mechanisms such as neuronal adaptation (e.g. Blake, 1989; Grossberg & Swaminathan, 2004; Wilson, 2007).

Here, we explore bistable perception in the DBM model. The DBM allows us to model bistability as originating from probabilistic inference in a hierarchical generative framework, where image interpretations are internally synthesised during the process of perceptual inference. Current research in cognitive science and computational neuroscience is concerned with the question of whether the brain might implement sampling-based approximate inference algorithms (e.g. Fiser et al., 2010). Boltzmann machines can use Gibbs sampling for inference (Section 2.1.3), which allows us to consider this issue of sampling in the brain. In particular, our work complements two recent models by Sundareswara & Schrater (2008) and Gershman et al. (2009b), who modelled perceptual bistability as sampling based-inference for the Necker cube and binocular rivalry, respectively. These models provide a high-level description specific to the bistability inference problem and corresponding algorithms used by the brain in this context. In contrast, our DBM model is intended as a model system of neuronal generative processing that is more generally applicable, as demonstrated in this thesis.

As was the case with the homeostatic adaptation underlying the emergence of hallucinations (Chapter 4), the fact that the DBM is a neural network makes it possible once again to explore concrete neuronal mechanisms. Here we address the question of the nature of the probabilistic sampling algorithm the brain might utilise. To that end, we provide a biological interpretation of a recently introduced sampling algorithm (Breuleux et al., 2011) as being based on neuronal adaptation (or neuronal fatigue and synaptic depression specifically), in close relationship to the homeostatic mechanism of Chapter 4 but on faster time-scales. Thus, we bridge the classic mechanistic and

the recent probabilistic explanations for bistability, suggesting that neuronal adaptation could actually be understood as a feature of a probabilistic sampling algorithm.

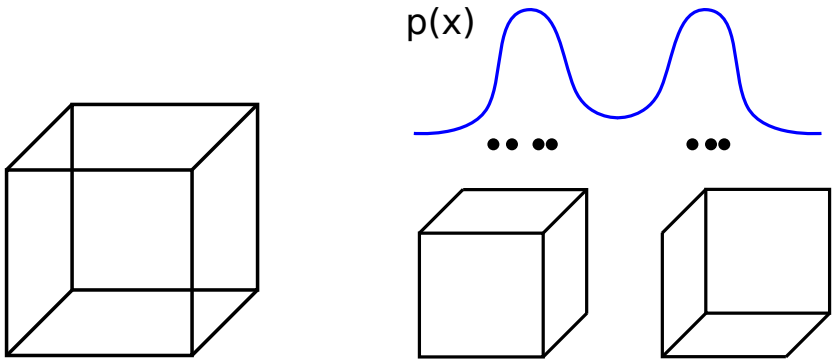
Our results focus on Necker cube bistability. Using a DBM that was trained on the two interpretations of the Necker cube, we show how adaptation leads to bistable switching of the internal percept when the model is presented with the actual ambiguous Necker cube. Moreover, we modelled the role of spatial attention in biasing the perceptual switching. Finally, we explored how the same approach can be applied to binocular rivalry as well. The majority of the presented results have been published (Reichert et al., 2011b).[2]

In the next section, we briefly review how perceptual bistability has been framed in the context of probabilistic inference in related computational studies. Subsequently, in Section 5.2, we delineate how the DBM can be used to model perceptual bistability. We describe the rates-FPCD sampling algorithm of Breuleux et al. (2011) and introduce its biological interpretation. In Section 5.3, we report on our results on modelling Necker cube bistability, including details of the simulation setups, analyses of the temporal dynamics of the network and of individual neurons, and a simulation involving spatial attention. Section 5.4 covers our simulation experiment concerned with binocular rivalry. Lastly, in Section 5.5, we discuss our overall results. In particular, we examine related computational models of probabilistic or Bayesian inference in bistability in more detail and contrast them to our own approach. We analyse how these approaches differ conceptually and in how they address the general issue of (approximate) probabilistic inference in the brain. We also discuss the role of noise and adaptation in bistability, as well as preliminary and potential future work on the topic in our modelling framework.

5.1 From probabilistic inference to bistable perception

Perception given ambiguous input can be framed as probabilistic inference under uncertainty. For the Necker cube (Figure 5.1), a possibly infinite number of configurations of lines in three dimensional space would be compatible with the observed 2D image, but a prior formed from experience of the world would make the two usually perceived wire-frame cubes the most likely interpretations in the posterior. To explain bistable perception in terms of probabilistic processing in the brain, it is necessary to relate

[2] In terms of content, the main changes are new and more extensive introduction and discussion sections, and some minor additions to the binocular rivalry experiment.



(a) Necker cube. (b) Bistability as probabilistic inference.

Figure 5.1: *(a)*: the ambiguous Necker cube. *(b)*: the two common perceptual interpretations of the Necker cube (bottom). Bistable perception might arise from probabilistic inference, exploring the posterior over explanations of the input over time. In a sampling-based representation, the posterior is approximated with a set of samples (top), one of which could correspond to the current conscious percept.

aspects of probabilistic representations and inference algorithms to features of bistable perception as observed in human subjects.

A first key aspect of perceptual bistability that needs to be addressed is how the inferred posterior over image interpretations given the image relates to the current percept as reported by the subject. In a full probabilistic model, the posterior would be a bimodal distribution. Apparently however, we seem to consciously perceive only one of the possibilities at a time,[3] which needs to be explained somehow by a probabilistic description of bistability. The second issue of bistability is then why the percept corresponding to one hypothesis or inferred mode of the posterior switches over time. Moreover, the dynamics of perceptual switching are also subject to stochasticity (e.g. Kang & Blake, 2010), which needs to be accounted for as well. To clarify, even though a probabilistic model deals with uncertainty or noise in the sensory input, the mapping from a fixed set of data to the posterior over some variables is in principle deterministic, as is variational approximate inference such as mean-field.[4] Stochasticity can come into play if posteriors are inferred over changing data samples subject to measurement noise (Stocker & Simoncelli, 2006). Alternatively, within a given probabilistic model,

[3] An exception might be conflicting images that would normally cause binocular rivalry but are perceived as summed at low contrast (Liu et al., 1992). Even in that case, one still does not perceive both images as concurrent *alternatives*.

[4] Variational inference uses optimisation, which itself might be stochastic (e.g. with stochastic gradient descent), but it will (if the learning rate is decreased appropriately) converge to a fixed result.

it could be the approximate inference algorithm used by the brain that is inherently stochastic.

5.1.1 Bistability and the sampling hypothesis

Several computational probabilistic models of perceptual bistability have been proposed (Dayan, 1998; Sundareswara & Schrater, 2008; Gershman et al., 2009b; Moreno-Bote et al., 2011). We review them below briefly to establish context for our work, and then consider them in detail in the discussion section of this chapter (Section 5.5). Generally, these models involve some form of approximate inference or approximate representation of probabilities to explain why only one image interpretation is perceived at a time. The model of binocular rivalry of Dayan (1998) also incorporates additional neuronal fatigue mechanisms, which account for deterministic switching but lack stochasticity and are not part of the probabilistic inference framework as such. The other aforementioned models on the other had base approximate inference on *sampling*, with the aim of explaining all aspects of bistability, including switching and stochasticity, within the probabilistic framework.

The notion that the brain uses a set of samples to approximate probability distributions has recently been explored from various perspectives theoretically and experimentally. In computational cognitive science, sampling algorithms have been suggested to connect idealised Bayesian models of cognition with the approximate inference people actually employ (Sanborn et al., 2010), basing their decisions on a small number of current hypotheses (i.e. samples from the posterior) that are updated in light of current data. This number of concurrently tracked hypotheses might be as low as one (Daw & Courville, 2008; Levy et al., 2009; Vul et al., 2009; Sanborn et al., 2010). On a lower level, neuroscience models posited concrete representational schemes in populations of neurons (Hoyer & Hyvärinen, 2003; Fiser et al., 2010). Here, one possibility is that the current state vector of a population would correspond to a single sampled point in high dimensional space. With neuronal firing rates, subject to noise, changing stochastically over time, several samples would be accumulated over multiple time steps. Fiser et al. (2010) also connected this hypothesis to data taken from early visual cortex.

When discussing his (non-stochastic) model of binocular rivalry, Dayan (1998) briefly mentioned a possible role for sampling. The idea that bistable perception could result from the brain using sampling to represent and explore the posterior is also put forward by Hoyer & Hyvärinen (2003), but they did not provide a computational model

of the phenomenon. The first concrete computational model to do so appears to be that of Sundareswara & Schrater (2008). There, Necker cube bistability is explained as resulting from the brain accumulating samples from the bimodal posterior over the orientation of the cube. To account for not only the stochasticity but also the concrete temporal dynamics of the perceptual switching as observed in human observers, Sundareswara & Schrater furthermore assume a perceptual decision process that explains both how individual samples are selected for awareness and how they decay over time to allow for switching to occur.

Modelling binocular rivalry, Gershman et al. (2009b) on the other hand proposed that specifically Markov chain Monte Carlo (MCMC) sampling could be well suited to explain perceptual bistability. As detailed in Section 2.1.3, MCMC methods produce samples from a distribution by iteratively exploring the state space. In particular, this generates correlated samples. Gershman et al. argue that an advantage of MCMC is that it is a general, "rational" approach to approximate inference in Bayesian models, out of which both the process of selection of the prevalent interpretation and the dynamics of perceptual switching fall out naturally: the current percept is identified with the current sample of the chain (which is in turn identified with the observer's current brain state), and the dynamics derive from the transition dynamics in the Markov chain. Hence, no additional ad-hoc selection process and memory mechanism are necessary.

As will be discussed in detail in Section 7.1, both Sundareswara & Schrater's and Gershman et al.'s sampling-based models of bistability are high-level descriptions, which are specific to the respective perceptual inference problems. They are not concerned at all with how this relates to neuronal mechanisms and representations, cortical processing, and the overall nature of a hypothetical probabilistic model of sensory input in the brain. Examining the latter issues is our aim here.

5.2 Neuronal adaptation in a DBM

We model perceptual bistability with the DBM model (Chapters 2 and 3). Because we use MCMC for inference in the DBM (specifically Gibbs sampling), our approach is following the proposal of Gershman et al. (2009b). Our goal is to examine the bistability phenomenon in a more general generative framework of cortical processing, which could account for both bistability from ambiguous images and from binocular rivalry, and do so based on internal representations *learnt* from data.

Moreover, we also explore the question of what of kind of biologically plausible

MCMC algorithm could be used by the brain, in particular one that goes beyond standard MCMC such as Gibbs sampling. It is well known that MCMC methods in general and Gibbs sampling in particular can be problematic in practice for complex, multimodal distributions, as sampling can get stuck in individual modes ('the chain does not mix', Section 2.1.3). To this end, we can yet again make use of the neural network nature of the DBM and give a neural interpretation to a recently introduced sampling algorithm that extends Gibbs sampling. In the process, we relate the classic neuronal adaptation accounts of bistability to the recent probabilistic models. This reasoning is explained in the following.

5.2.1 The rates-FPCD sampling algorithm

The sampling algorithm we employ in this study was recently proposed for restricted Boltzmann machines (RBMs, Section 2.1.4) by Breuleux et al. (2011). It is called rates-FPCD (Rates Fast Persistent Contrastive Divergence), and is an adaptation of the training algorithm FPCD, which in turn is based on PCD, as was introduced briefly in Section 2.2.1.

To recapitulate, PCD is one of the methods used to approximately compute the likelihood gradient for the weight update when learning RBMs. It approximates the negative phase terms, i.e. the expectations over the model distribution $\langle v_i h_j \rangle_{P(\mathbf{h},\mathbf{v})}$ (where $\mathbf{v}$ and $\mathbf{h}$ are the states of the visible and hidden units, respectively), by having the model freely generate 'fantasy' samples in a Markov chain that is persistent across weight updates (see e.g. Figure 2.3 on page 34). If the learning rate and weight updates are very small and the chain mixes, then the produced samples across several weight updates will faithfully represent the model distribution at that point during learning (Tieleman, 2008).

In practice however it turns out that PCD is effective with a higher learning rate than expected. As Tieleman & Hinton (2009) show, this is because updating the model parameters with the negative phase terms computed with PCD has an unintended side-effect of improving the mixing rate of the persistent chain. If mixing is slow, the fantasy particles can get stuck in a high-probability mode of the current model distribution. The negative phase update however has the effect of increasing the energy of the states currently occupied by the particles, thus decreasing the height of this mode under model. Across several weight updates, this will reduce the probability of the particles to remain in this mode, essentially pushing them out of the energy valley they were stuck in by

reducing the latter's depth. This thus can improve mixing, though it has the disadvantage of changing the model distribution in a somewhat uncontrolled way. For example, the mode the particles were stuck in could be one that is actually supported by the data, so decreasing probability mass there disproportionally is not actually desirable and then has to be counter-acted in the data-driven positive phase.

Tieleman & Hinton (2009) capitalise on this effect by introducing an additional set of 'fast weights' next to the regular model parameters, yielding the Fast PCD (FPCD) algorithm. These fast weights are updated from the gradient similarly to the regular parameters, but they are only actually used in the negative phase, adding them to the regular parameters there in order to improve the mixing. They change with a relatively large learning rate but also decay faster. In essence, the fast weights add a temporary overlay on the energy landscape (as it is defined by the regular parameters) to improve mixing in the negative phase.

Concretely, with the visible layer in the RBM indexed by 0 and the hidden layer by 1, the fast weights $\mathbf{W}_f$ and the analogue fast biases $\mathbf{b}_f^{(k)}$ are updated during a gradient step according to

$$\mathbf{W}_f \leftarrow \alpha\mathbf{W}_f + \varepsilon(\mathbf{x}^{(0)} p(\mathbf{x}^{(1)}|\mathbf{x}^0) - \mathbf{x}'^{(0)}\mathbf{x}'^{(1)T}), \tag{5.1}$$

$$\mathbf{b}_f^{(0)} \leftarrow \alpha\mathbf{b}_f^{(0)} + \varepsilon(\mathbf{x}^{(0)} - \mathbf{x}'^{(0)}), \tag{5.2}$$

$$\mathbf{b}_f^{(1)} \leftarrow \alpha\mathbf{b}_f^{(1)} + \varepsilon(p(\mathbf{x}^{(1)}|\mathbf{x}^0) - \mathbf{x}'^{(1)}). \tag{5.3}$$

The terms in the parentheses correspond to the positive and negative phases. The visible states $\mathbf{x}^{(0)}$ are clamped to the current data item,[5] whereas $\mathbf{x}'^{(0)}$ and $\mathbf{x}'^{(1)}$ are current samples from the model run freely in the persistent chain. ε is a parameter determining the rate of adaptation, and $\alpha \leq 1$ is a decay parameter that also limits the amount of weight change contributed by the fast weights. The second term in each of the parentheses has the effect of changing the weights and biases such that whatever states are currently being sampled by the model are made less likely in the following. Hence, this will eventually 'push' the chain out of a mode it is stuck in. The first terms in the parentheses are computed over the data and lead to the chain being drawn to states supported by the current input.

FPCD is used during RBM training. Breuleux et al. (2011) adapted FPCD to turn it into a general algorithm for producing samples from a RBM applicable outside of learning, intending to make use of the improved mixing due to the temporarily changing fast weights (whereas the regular model parameters now remain fixed). The

[5] In practice, minibatches are used.

key difference is that outside training, the training data is not necessarily available anymore, hence the data terms depending on $\mathbf{x}^{(0)}$ in the equations above need to be modified. Breuleux et al. made a simple (heuristic) change and replaced the pairwise and unitary statistics calculated from current data with averages computed once over *all* training data:

$$\mathbf{W}_f \leftarrow \alpha\mathbf{W}_f + \varepsilon(E[\mathbf{x}^{(0)}\mathbf{x}^{(1)T}] - \mathbf{x}'^{(0)}\mathbf{x}'^{(1)T}), \tag{5.4}$$

$$\mathbf{b}_f^{(0)} \leftarrow \alpha\mathbf{b}_f^{(0)} + \varepsilon(E[\mathbf{x}^{(0)}] - \mathbf{x}'^{(0)}), \tag{5.5}$$

$$\mathbf{b}_f^{(1)} \leftarrow \alpha\mathbf{b}_f^{(1)} + \varepsilon(E[\mathbf{x}^{(1)}] - \mathbf{x}'^{(1)}) \tag{5.6}$$

The authors denote these statistics as *rates*, hence the name rates-FPCD. These quantities need to be computed once, doing inference at the end of training with all data still available, but can from then used for sampling from the model while the training data can be discarded. It was found that these terms serve to sufficiently stabilise the sampling scheme, and overall that rates-FPCD yielded improved performance over standard Gibbs sampling.

5.2.2 Biological interpretation

Let us consider Equations 5.4-5.6 from a biological perspective, interpreting the weight parameters as synaptic strengths and the biases as overall excitability levels of the neurons. The algorithm suggests that the capability of the network to explore the state space is improved by dynamically adapting the neuron's parameters (cf. e.g. Maass & Zador, 1999). The adaptation depends on the current states of the neuron and its connected partners (second terms in parentheses), drawing them towards some set values (first terms, the rate statistics). To implement the adaptation, a neuron needs to remember its average firing activity at the end of learning (for the bias statistics), and the synapses an average firing correlation between connected neurons (for the weight statistics). The mechanism underlying rates-FPCD can thus be interpreted as neuronal adaptation. While Breuleux et al. (2011) did not make any appeal to biology when introducing rates-FPCD, Welling (2009) did suggest in closely related work (on dynamic weights for sampling) a connection to dynamic synapses in the brain.

For our study here, we use this neuronal adaptation for a model of perceptual bistability as is caused by ambiguous or conflicting sensory input. Unlike Breuleux et al., we thus apply rates-FPCD not to sample from the full model *joint* distribution (which includes sampling the visible layer), but rather to sample from the *posterior* during

perceptual inference given an input image (hence the visible layer remains clamped to that image). We also use a DBM, rather than just a RBM, and thus extend the adaptation to all sets of weights and biases across multiple hidden layers.

The neuronal adaptation as defined by Equations 5.4-5.6 is closely related to the homeostatic mechanism we introduced in Chapter 4, though it also acts on the weights and not just the biases. The key differences however are the time-scales involved and the functional roles of the mechanisms. The homeostatic adaptation of Chapter 4 was meant to model longer term regulation of firing rate over hours or days. An adaptation step was thus applied only after averaging current activity levels over many trials involving many input images. Rates-FPCD as utilised by us on the other hand adapts the network parameters during a single trial given a single input image, and thus is better interpreted to operate on timescales in the order of seconds, dynamically affecting the outcome of ongoing perception. Hence, while the functional role of homeostasis for perception was a stabilisation of the *overall* internal representations, e.g. to counteract input degradation, the neuronal adaptation of rates-FPCD serves to improve fast exploration of the currently inferred posterior.

As a final note, Equations 5.4-5.6 can be made to implement a specific kind of neuronal adaptation, namely neuronal fatigue and synaptic depression. If activation patterns in the network are learnt to be sparse and neurons are thus off most of the time, then the resulting rate statistics are of low values, and a neuron will fire strongly only for its preferred stimulus (or stimulus interpretation).[6] Neuronal adaptation back to baseline then corresponds to an excitability and synaptic efficacy drop. The resulting changes allow the network to discover potential alternative interpretations of the stimulus.

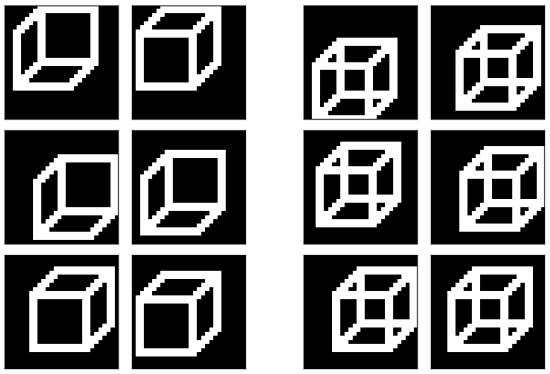
5.3 Experiments: the Necker cube

We sought to model perceptual bistability with the DBM, using the sampling algorithm rates-FPCD, which could be interpreted as implementing neuronal adaptation. In our first set of experiments, we used the ambiguous Necker cube image as input.

5.3.1 Methods and model setup

According to the probabilistic account, bistable perception is a result of the brain actively exploring different hypotheses or image interpretations that could explain the

[6] The role of sparsity for the functional role of neuronal representations will be discussed in Section 6.4 in the chapter on attention.



(a) (b)

Figure 5.2: *(a)*: examples of the unambiguous images used in training. *(b)*: example images of the ambiguous Necker cube.

sensory data. To obtain a DBM that could infer the two interpretations of the Necker cube, we trained a DBM on binary images of cubes at various locations, representing the two unambiguous interpretations of the Necker cube (Figure 5.2a). The underlying assumption is that when the brain encounters such objects in the world and learns to represent them, they are unambiguous most of the time. The experiments then tested the model on images of the actual, ambiguous Necker cube (Figure 5.2b).

The overall model setup was similar to that described in the other parts of this thesis. The DBM had a visible layer with 28×28 units corresponding to the size of the training images, and three hidden layers with 26×26 units each. Receptive fields in the weights were restricted, with sizes 7×7, 13×13, and 26×26 from lowest to highest hidden layer. The images in the training data set covered all possible positions of the cubes, so there was no generalisation involved apart from that unambiguous cubes were then replaced with Necker cubes. The total number of images was 60,000 (simply for consistency with the earlier experiments in this thesis, Section 4.2.2). Layer-wise training used CD-1 (30 epochs; see Section 3.6.1 for further training parameters) and no further learning was performed once the DBM was composed. The rate statistics were computed by measuring unit activities and pairwise correlations from running the model on the training data after training was completed.

As before, we decoded the hidden states (Section 3.5) to examine what the internal state of the model represented during perceptual inference. When neuronal adaptation was used, the adaptation rate parameter was set to $\varepsilon = 0.001$, and the decay parameter to $\alpha = 0.95$. These values were chosen to lie in a range where bistability was caused by ambiguous input, while internal representations were still stable for unambiguous images. In general throughout this work, a trial consisted of clamping the visible layer to a randomly selected image from the relevant data set and running inference for a

variable number of steps. Unless noted otherwise, the hidden states were initialised to zeros at the beginning of each trial.

5.3.2 Bistable perception from neuronal adaptation

For a first initial experiment, we tested the model on the Necker cube, after it had been trained on unambiguous interpretations, using only normal Gibbs sampling to sample the hidden states and no rates-FPCD/neuronal adaptation. The hidden states were initialised to zero and the visible layer clamped to one of the Necker cube images. Over several repeated experiments, when sampling the hidden units the decoded hidden states were then found to converge within a few sampling cycles to one of the unambiguous interpretations and remained therein, exhibiting no perceptual switching to the respective alternative interpretation. This does not mean however that the latter did not constitute a stable mode in the inferred posterior: if the hidden states were initialised to the alternative interpretation (e.g. by first doing inference over the respective unambiguous image and then exchanging the input for the Necker cube), the internal representations maintained the latter as well. Thus, both interpretations were in principle captured by modes in the posterior and were stably represented by the model if found, but Gibbs sampling displayed bad mixing and failed to explore the distribution sufficiently.

It should be noted that, generally speaking, the behaviour of the network strongly depends not only on the sampling algorithm used during inference but also on the model parameters learnt in training. We employed only the most simple training methods, namely layer-wise pre-training with CD-1 and no tuning of the full DBM. It is known that models trained with CD-1 suffer from problems related to hidden units not exploring the state space (Section 2.2.1, Desjardins et al., 2010; Breuleux et al., 2011). Thus we here do not want to imply that, using better training, it would not be possible to learn a DBM that mixes better even with Gibbs sampling, especially for this simple data set.[7] However, for the argument at hand it is more important that in general, bad mixing can be a problem that might be alleviated by methods such as rates-FPCD, or neuronal adaption in the biological sense, hence using a setup that exhibits this problem is useful to make the point.

We then employed rates-FPCD to model neuronal adaptation and repeated the ex-

[7] Indeed, preliminary experiments using models trained with PCD rather than CD resulted in some spontaneous switching. However, without further fine-tuning of the learning parameters, reconstructions were noisy and the two modes less stable, thus these models appeared less useful for the study of bistability and we did not explore that avenue further.

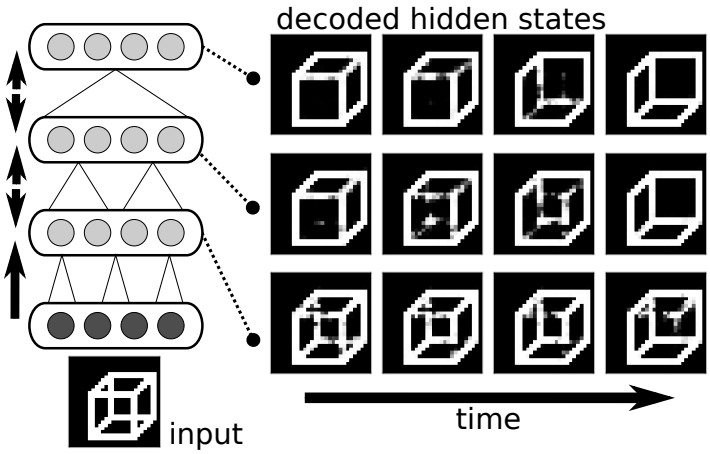


Figure 5.3: Example decoded states exhibiting bistable perception. During inference on an ambiguous image, the decoded hidden states reveal perceptual switching resulting from neuronal adaptation. Four consecutive sampling cycles during a transition are shown.

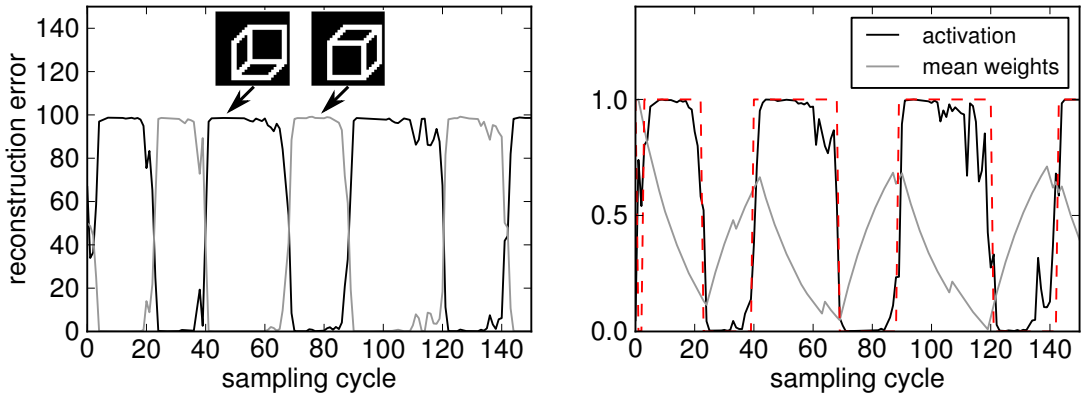
periment. With neuronal adaption, the internal representations as decoded from the hidden layers were found to switch over time between the two image interpretations. Thus the model exhibited perceptual bistability. An example of the switching of internal representations is displayed in Figure 5.3. It can be observed that the perceptual state is most distinct in higher layers.

5.3.3 Relation between perceptual state and individual neurons

For quantitative analysis, we computed the squared reconstruction error of the image decoded from the topmost layer with regards to either of the two image interpretations. Plotted against time (Figure 5.4a), this shows how the internal representations evolved during a trial. The representations matched one of the two image interpretations in a relatively stable manner over several sampling cycles, with some degradation before and a short transition phase during a perceptual switch.

To examine the effects of adaptation on an individual neuron, we picked a unit in the top layer that showed high variance in both its activity levels and neuronal parameters as they changed over the trial, indicating that this unit was involved in coding for one but not the other image interpretation. In Figure 5.4b are plotted the time course of the neuron's activity levels (i.e. firing probability) and the mean synaptic efficacy, i.e. weight strength, of connections to this unit.[8]

[8]The changes to weights and biases were equivalent, so we show only the former.



(a) Match of internal state to interpretations.

(b) Single unit properties.

Figure 5.4: *(a)*: time course of squared reconstruction errors of the decoded topmost hidden states, w.r.t. either of the two image interpretations (dark and light curves, respectively). Apart from during the transition periods, the percept at any point in time matched one (close to zero error) but not the other interpretation (high error). *(b)*: activation (i.e. firing probability, dark curve) and mean synaptic strength (light curve, arbitrary origin and units) of a top layer unit that participated in coding for one but not the other interpretation (dashed line marks currently active interpretation). Synaptic efficacy depressed during instantiation of the preferred interpretation and recovered during the non-preferred interpretation. This neuronal adaptation caused changes in activation leading up to each subsequent perceptual switch.

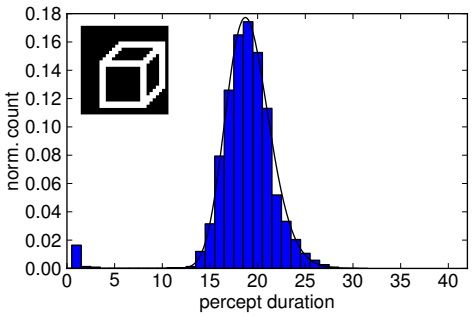
As expected, the firing probability of this unit was close to one when one of the interpretations was currently instantiated, and close to zero for the other. Especially in the initial time period following a perceptual switch, the activation was stable at high or low levels. However, as the neuron's firing rate and synaptic activity deviated from their low average levels captured by the rate statistics, the synaptic efficacy changed, as shown in the plot. For example, during instantiation of the preferred stimulus interpretation, a drop of neuronal excitability ultimately lead to a waning of activity that preceded and, together with the changes in the overall network, subsequently triggered the next perceptual switch.

In another trial, we analysed the same neuron as before but with an input image where the Necker cube was in a different position. The neuron then showed constant low firing rates, indicating that it was not involved in representing that image. The neuronal parameters were found to be stable throughout the trial, after a slight initial monotonic change that would allow the neuron to assume its low baseline activity as determined by the rate statistics.

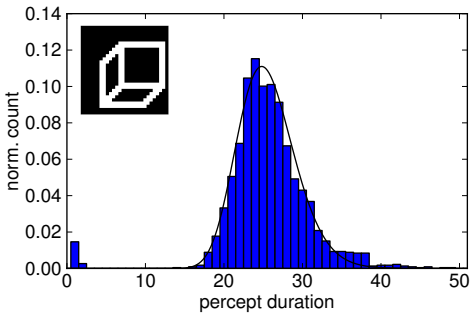
Moreover, other units were found to have relatively stable high firing rate for a given image throughout whole the trial, coding for features of the stimulus that were common to both image interpretations. The high firing rate was maintained even though these neurons' parameters equally adapted due to elevated activity. This is due to the extent of adaptation being limited by the decay parameter α (Equations 5.4-5.6). It shows that the adaptation can be set to be sufficiently strong to allow for exploration of the posterior, without overwhelming the representation of unambiguous image features. Similarly, as noted in the methods section, internal representations of the model when presented with the *un*ambiguous images from the training set were stable under adaptation with the same setting of adaptation parameter values.

5.3.4 Temporal statistics of bistability

We also quantified the statistics of perceptual switching by measuring the length of time the model's state would stay in either of the two interpretations. For a prolonged trial over a single input image, the resulting histograms of percept durations, i.e. time intervals between switches, are displayed in Figure 5.5, separately for the two interpretations of the image. The histograms are shaped like gamma or log-normal distributions (plots show log-normal fits), qualitatively in agreement with experimental results in human subjects (Zhou et al., 2004). Because we did not expect a quantitative match to such ex-



(a) Percept durations for interpretation one.



(b) Percept durations for interpretation two.

Figure 5.5: *(a) + (b)*: histograms over percept durations between perceptual switches during one trial, for both interpretations (insets) of the test image, respectively. Ignoring the peaks at small interval lengths, which stem from fluctuations during transitions, the histograms are well fitted by log-normal distributions (black curves).

perimental data (given that the implementation specifics of the DBM are rather different from those of real brains, and given quantitative and qualitative variation across human subjects, op. cit.), we did not originally attempt a quantitative comparison. Still, a more detailed analysis, including of the influence of model parameters on the statistics, might be beneficial.

What would also complement our empirical results here would be a principled theoretical examination of the relation between the rates-FPCD algorithm and the observed temporal dynamics. In particular, due to the ongoing adaptation of neuronal parameters, the resulting walk in the state space will no longer be Markovian, i.e. the probability of the next state will not only depend only on the current one. The latter is the case in standard Gibbs sampling. Thus, rates-FPCD should introduce correlations between subsequent percept durations. In fact, Zhou et al. (2004) argued that experimentally observed percept durations are better described by log-normal than gamma distributions, and that this suggests exactly such correlations. Similarly, analysis of experimental data and theoretical work on spike trains of individual neurons shows that intervals between spikes are correlated and, at least in some cases, well described by log-normal distributions, which could be a result of neuronal adaptation (Farkhooi et al., 2009). It might thus be possible to establish a theoretical connection between the dynamics of individual neurons and the switching dynamics of the network as a whole, or to at least utilise analysis methods applied to the former for the latter. Due to time being limited, we could not explore this avenue so far. We will briefly address the related issue of the

interaction between stochasticity and adaptation in the discussion (Section 5.5.3), there in the context of reproducing a relevant psychophysics experiment.

As can be seen by comparing the histograms in Figure 5.5, there was a bias apparent in the model towards one of the interpretations (found to be different for different images). Some biases are observed in humans (as visible for instance in the data of Sundareswara & Schrater, 2008), potentially induced by statistical properties of the environment. However, our training data set did not involve any biases, so this seems to be merely an artifact produced by the (basic) training procedure used. Finally, in the histograms there are also small peaks at very short percept durations. These result from fluctuations of the current percept and its classification according to the two alternative interpretations during brief phases of transitioning from one percept to the next (see e.g. the decoded states in Figure 5.3, as well as the time course of the percept in Figure 5.4a). In reality, such brief transition phases might not be perceptually salient, though through introspection it appears to us that the switching of the Necker cube percept indeed appears to be gradual and not instantaneous.

5.3.5 The role of spatial attention

The statistics of multistable perception can be influenced voluntarily by human subjects (Meng & Tong, 2004). For the Necker cube, overtly directing one's gaze to corners of the cube, especially the interior ones, can have a biasing effect (Toppino, 2003). This could be explained by these features being in some way more salient for either of the two interpretations. An explanation matching our (simplified) setup would be that opaque cubes (as used in training) uniquely match one of the interpretations and lack one of the two interior corners. In the following, we model not eye movements but covert attention, involving only the shifting of an internal attentional 'spotlight', which also has been shown to affect perceptual switching in the Necker cube (Peterson & Gibson, 1991).

In this experiment, the presented image remained unchanged, but a spatial spotlight that biased the internal representations of the model was employed in the first hidden layer. To implement the spotlight, we made use of the fact that receptive fields were topographically organised, and that sparsity in a DBM breaks the symmetry between units being on and off and makes it possible to suppress represented information by suppressing the activity of specific hidden units (this attentional mechanism will be discussed in detail in Chapter 6). We used a Gaussian shaped spotlight that was centred

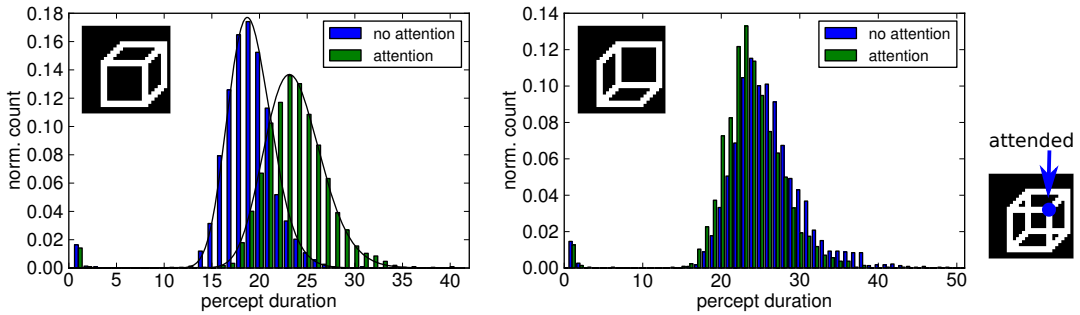
at one of the salient internal corners of the Necker cube (Figure 5.6c) and applied it to the hidden units as additional negative biases, attenuating activity further away from the focus.

The effect of attention on the percept durations for one of the test images is displayed in Figure 5.6, together with the data obtained earlier without attention for comparison. For the interpretation that matched the corner that was attended, we found a shift towards longer percept durations (Figure 5.6a). On the other hand, the distribution for the other interpretation was relatively unchanged (Figure 5.6b). This effect was consistent across the test data. Averaged over all test images, the mean interval spent representing the interpretation favoured by spatial attention saw a 25% increase, whereas there was approximately no change for the other interpretation. Hence, in the model spatial attention prolonged the percept whose salient feature is being attended.

This is qualitatively in line with experimental data at least in so far that voluntary attention does indeed have an effect. We did not find an experimental study matching exactly our simulation experiment, namely examining covert attention on the interior corners in unmodified Necker cubes. For example, Peterson & Gibson (1991) considered the role of spatial attention but with versions of the Necker cube that contained regions biased towards one of the interpretations. Other studies that include Necker cube bistability focus on attentional control not specified to be spatial attention (e.g. Meng & Tong, 2004; van Ee et al., 2006) and/or on overt spatial attention mediated by eye movements (e.g. Toppino, 2003). These differences aside, experimental findings generally do not appear to confirm our result where only the attended percept is affected whereas the other remains mostly unchanged. Notably, van Ee et al. (2006) found that subjects, when explicitly instructed to do so, could selectively prolong only one of the interpretations while leaving the distribution of the alternative relatively unchanged. However, it is not clear that this instruction would correspond to our simulation where a spatial spotlight was present throughout the whole trial. Further simulations and analysis might be necessary to explain our results and a possible discrepancy with experiments, as might be performing a psychophysics experiment matching our simulation setup.

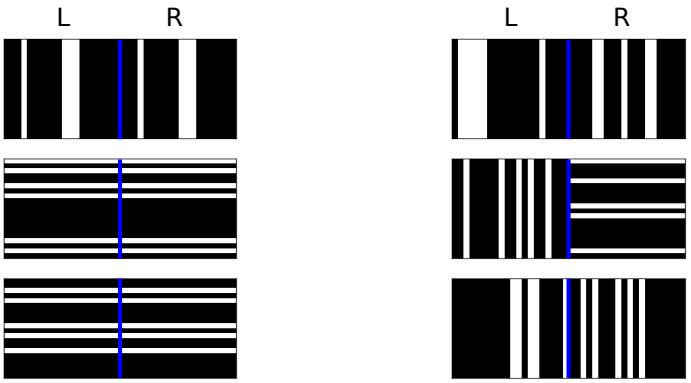
5.4 Experiments: binocular rivalry

Several related studies that considered perceptual multistability in the light of probabilistic inference focused on binocular rivalry (Dayan, 1998; Hohwy et al., 2008; Gershman et al., 2009a). There, human observers are presented with a different image to each eye,



(a) Percept durations for interpretation one. (b) Percept durations for interpretation two. (c)

Figure 5.6: *(a) + (b)*: histograms over percept durations for both interpretations (insets) of the test image, respectively, both with or without spatial attention. Black curves are log-normal fits (omitted in (b) to avoid clutter). Spatial attention was employed to the top-right interior corner of the Necker cube (as shown in *(c)*). Note that spatial attention acts on the hidden states, rather than directly on the input image in the visible layer, as explained in the text. The attended corner was salient for the first interpretation from (a) because for this feature, ambiguous input image and internal interpretation matched. Attention lead to a shift in the distribution towards longer percepts. For the second interpretation from (b), the attended corner in the input conflicted with the interpretation. The percept durations for that interpretation were unchanged.



(a) Training images. (b) Test images.

Figure 5.7: Example images for the binocular rivalry experiment. *(a)*: training images contained either horizontal or vertical bars, and the left and right image halves were identical (corresponding to the left and right eyes). *(b)*: For the test images, the left and right halves were drawn independently. They could come from the same category (top and bottom examples) or from conflicting categories (middle example).

and their perception is found to switch between the two images. Depending on specifics such as size and content of the images, perception can switch completely between the two images, fuse them, or do either to varying degrees over time (Tong et al., 2006; Knapen et al., 2007). We demonstrate with a simple experiment that the phenomenon of binocular rivalry can be addressed in our framework as well.

To this end, the same model architecture as before was used, but the number of visible units was doubled and the units were separated into left and right 'eyes'. During training, both sets simply received the same images. During testing however, the left and right halves were set to independently drawn training images to simulate the binocular rivalry experiment. The units in the first hidden layer were constrained to be monocular in the sense that their receptive fields covered visible units only in either the left or right half. Higher layers did not make this distinction. As a data set we used images containing either vertical or horizontal bars (Figure 5.7), similar to the images used in the 'bars problem' (Földiák, 1990; Spratling, 2011). Unlike in the latter, here only either vertical or horizontal bars were present in any one image, resulting in two separate image categories (which will become relevant below). Adaptation parameters in this experiment were set to $\alpha = 0.9$, $\varepsilon = 0.002$. Model parameters and training procedure were otherwise as in the Necker cube experiments.

As with the Necker cube, perceptual switching between competing images was

observed with neuronal adaptation but not without. At the same time, for identical images as used in training, percepts were mostly stable even with adaptation with our choice of adaptation parameter values, apart from occasional flickering of individual bars in the percept.

The model setup and example decoded states in a trial are shown in Figure 5.8. Generally, the perceptual state was found to be biased to one of the two images for some periods, while fusing the images to some extent during transition phases (Figure 5.9). Interestingly, whether fusing or alternation was more prominent depended on the nature of the conflict in the two input images: for images from the same category (both vertical *or* horizontal lines), fusing occurred more often (Figure 5.9a), whereas for images from conflicting categories, the percept represented more distinctly either image and fusing happened primarily in transition periods (Figure 5.9b).

To quantify this, we computed the reconstruction errors from the decoded hidden states with regards to the two images, and took the absolute difference averaged over the trial as measure for how much the internal states were representing both images individually rather than fused versions. Averaged over all trials, this measure was more than two times higher for input images from conflicting categories than for those from compatible ones. This result is qualitatively in line with psychophysical experiments that showed more fusing when images were different but compatible (e.g. different patches of the same source image, Tong et al., 2006; Knapen et al., 2007).

Finally, a long-standing issue in the research on binocular rivalry is where in the visual system the neuronal activity reflects the currently active percept, and whether the rivalry is mostly a result of competition between eyes (interocular competition) as represented by monocular neurons, or between image content (pattern competition) as represented by binocular neurons higher in the cortical hierarchy (Tong et al., 2006; Leopold & Logothetis, 1999). Current evidence favours a gradual change, with higher cortical areas such as IT strongly modulated by the current percept and lower areas like V1 weakly, as well as a hybrid model that sees inhibitory interactions between neuronal populations at many cortical sites (Tong et al., 2006). There is some disagreement on the extent of the involvement of early stages such as V1 or even LGN, with neurophysiological studies finding much weaker effects of rivalry than neuroimaging ones.

In our model, we found a similar gradient along the hierarchy, as can be seen in the example decoded states in Figure 5.8. The neurons in the first, monocular hidden layer were driven mostly by their respective input images, meaning that the first layer

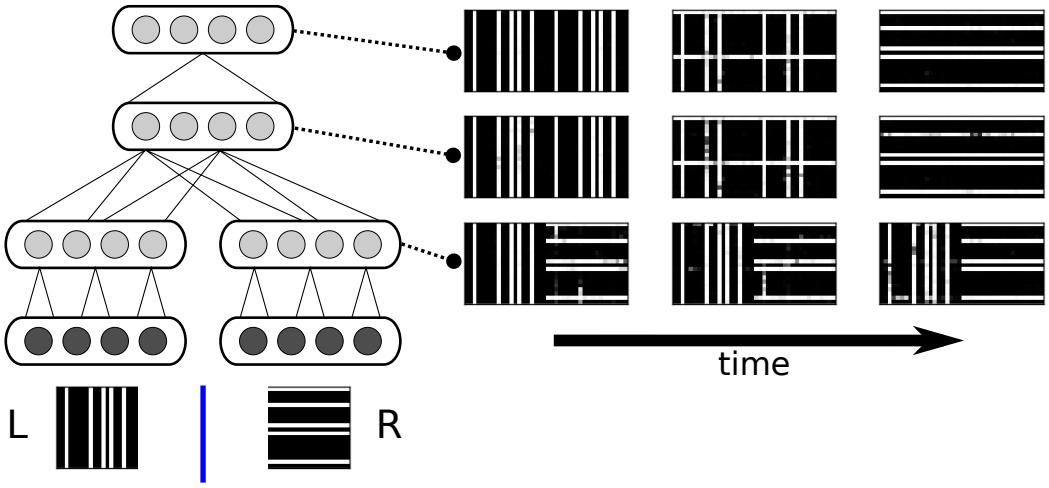
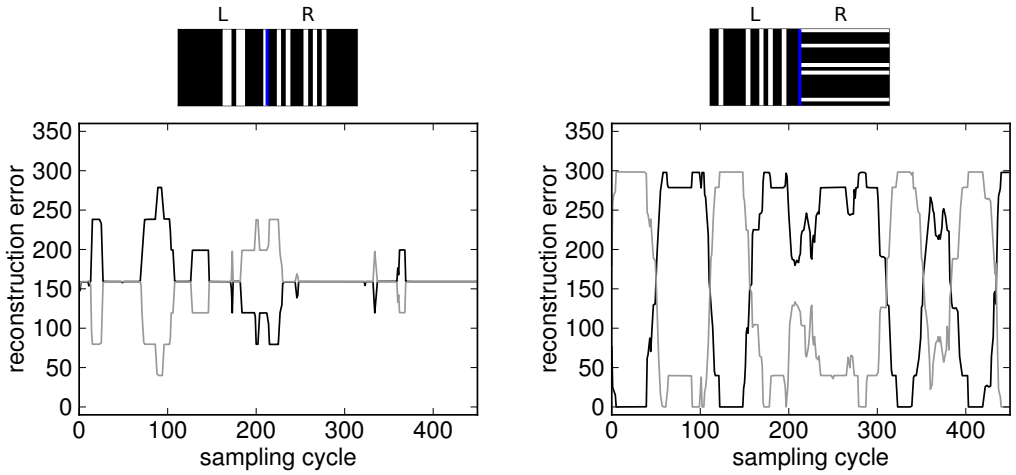


Figure 5.8: Model setup in the binocular rivalry experiment (left) and example decoded states during perceptual bistability (right). For this experiment, the visible units were doubled and separated into two halves corresponding to left and right eye, and the model was trained on images identical in both. The units in the first hidden layer were monocular. To evoke binocular rivalry, different images were shown to left and right eyes (bottom left). Shown are the decoded hidden states at three points in time 25 sampling cycles apart. The percept switched between the alternatives over time, but there were also periods where the percept fused them in various ways (middle column). Across the hidden layers, the higher two expressed the current percept strongly. The neurons in the first, monocular hidden layer mostly represented the image from the respective eyes. However, even in the first layer there is some interference visible for whichever image percept that was currently suppressed in the higher layers. This results from feedback from the latter.



(a) Percept vs. eye images for same category.

(b) Percept vs. eye images for conflict. categories.

Figure 5.9: For one trial of the binocular rivalry simulation, displayed are the squared reconstruction errors for decoded top layer representations computed against either of the two input images. *(a)*: the input images came from the same category (here, vertical bars), and fusing of the percept was prominent, resulting in modest, similar errors for both images. *(b)*: for input images from conflicting categories, the percept alternated more strongly between the images, although intermediate, fused states were still more prevalent than was the case for the Necker cube. The step-like changes in the error were found to result from individual bars appearing and disappearing in the percept.

overall represented both the two conflicting images. On the other hand, the other two hidden layers strongly expressed a percept that alternated over time.[9] This change in representations was rather abrupt between the first and second hidden layer. With additional intermediate hidden layers this might have been more gradual, as might be supported by electrophysiological data from different cortical areas along the hierarchy (Leopold & Logothetis, 1999), but we did not test this.

At the same time, in the decoded examples there is some effect of rivalry visible even in the first hidden layer, with some interference to the image half that was at that time suppressed in the higher layers. Because input to the monocular neurons in the first hidden layer is from either half only, this interference is necessarily an effect of feedback from the second hidden layer. It should be noted that, because there are no lateral connections within a layer in the DBM, it is not really possible to differentiate between lateral interactions and feedback from higher areas in the model.

To find out whether an interaction with monocular neurons was necessary to evoke rivalry, in another experiment we drove the first layer with feedforward input only (using the acetylcholine mechanism, Section 4.3.6). Representations in the first layer then faithfully represented both the two conflicting images without interference. Nevertheless, higher hidden layers still exhibited perceptual bistability. Thus, in the model at least, feedback can lead to an effect of rivalry on monocular neurons, but no changes in the monocular representations are necessary to realise rivalry, nor are interactions between the monocular neurons, be it indirectly via feedback or directly via possible lateral connections. Hence, in the model bistability would be better described as resulting from pattern competition rather than interocular competition (Tong et al., 2006; Leopold & Logothetis, 1999).

As mentioned above, there is a discrepancy between findings from neuroimaging and electrophysiological studies regarding how extensively perceptual rivalry is realised in early stages of processing in the visual system. A possible explanation could be that fMRI might correlate more with synaptic activity than neuronal spiking, and that measured signals in early stages showing correlation with the current percept thus reflect feedback input from higher areas more than actual local spiking activity (Tong et al., 2006). In the model, this is mirrored by representations in the first hidden layer that

[9] It should be noted that, because the decoding procedure (Section 3.5) maps the hidden states back into the full input space going through the lower hidden layers, the decoded images always consist of two input images corresponding to the two eyes. It can be observed in Figure 5.8 that the higher hidden layers had indeed learned binocular representations, which meant that the decoded images were nearly identical in both halves when decoded back into input space. This was the case even when the percept was actually a fusion of the images from both eyes (middle column in the figure).

were only weakly modulated by rivalry, despite the next higher hidden layer expressing rivalry strongly and thus sending feedback to the first that was correlated with the alternating percept.

5.5 Discussion

We modelled bistable perception as arising from sampling-based probabilistic inference, and in this context provided a biological interpretation to the rates-FPCD sampling algorithm in terms of neuronal adaptation. Compared to other computational models of bistable perception, our model uniquely combines the following aspects: the DBM is not meant as theoretical description of specifically perceptual bistability, but rather as a versatile model system of various aspects of processing in a generative framework (Chapters 4 and 6); as a neural network, it thus uses learned, distributed representations, rather than a hard-wired neuronal circuitry, and can be applied to both bistability from the ambiguous Necker cube and from binocular rivalry; it allows for relating neuronal adaptation to sampling-based probabilistic inference (specifically using MCMC), thus giving the former a normative context; it implements a spatial attention mechanism; and finally, it involves *both* noise and adaptation, as has recently been argued to be necessary to account for the properties of bistable perception (Shpiro et al., 2009; Kang & Blake, 2010).

As stated in the first chapter, one of our goals for this thesis is to examine approximate probabilistic inference in the brain. In the following, we consider related probabilistic models of bistability in more detail, contrasting them to our own model. To this end, we make use of our terminology for characterising Bayesian models as was introduced at the beginning of this thesis (Section 1.2.3). Given the current debate concerning Bayesian approaches to modelling cognition (Section 1.2), we also aim to demonstrate with this more detailed discussion how different notions of Bayesian models can be disentangled. Furthermore, we then address the question of how different approximate inference schemes relate to perceptual experience. Lastly, we elaborate on remaining issues surrounding our approach to modelling bistable perception, including the interplay of noise and adaptation, preliminary work on including depth information to model Necker cube bistability, and possible future work.

5.5.1 Related work: an analysis of Bayesian models of perceptual bistability

We discuss in detail four probabilistic models of bistability (Sundareswara & Schrater, 2008; Gershman et al., 2009b; Moreno-Bote et al., 2011; Dayan, 1998).[10] To guide the analysis, we differentiate (see Section 1.2.3 for explanation) between *high-level* (psychological) and *low-level* (neuronal), and *external* (describing the world) and *internal* (describing the brain) models. We also distinguish between *conceptual* models with latent variables that have by design a clear meaning or interpretation (such as, the orientation of a hypothetical cube underlying the Necker cube image), and *instrumental* models, which primarily serve a purpose (such as, forming a useful latent representation of sensory data), with no explicit meaning assigned to the involved latent variables *a priori*.

A high-level account of the Necker cube problem

Sundareswara & Schrater (2008) model bistability from ambiguous images (the Necker cube) as arising from approximate inference in a Bayesian model, using a custom sampling procedure. They focus on a quantitative analysis of the switching dynamics and how it matches human observers, and the influence of image context. Underlying their approach is an explicit description of the perceptual problem as inferring a variable capturing the orientation of the cube from the data. The authors give a high-level description of the inference process leading to bistability. They pose that the brain produces i.i.d. samples from the bimodal posterior and accumulates them over time. Each sample is assigned a weight that is determined from the height of the posterior at the corresponding point and also, importantly, from a discount factor exponential in the age of the sample, implementing some form of memory decay. Only one sample however is 'chosen' as conscious percept, namely the one with maximal weight. The results of their study thus crucially depend on this custom selection and memory process, which is not motivated further by the authors either on normative or neurological grounds.

Sundareswara & Schrater's discussion generally suggests an *internal* interpretation of their model (in particular, they do not use the terms 'optimal', 'ideal observer', or

[10] For comparison, the work by Grossberg & Swaminathan (2004) is an example of a general but non-probabilistic framework addressing Necker cube bistability. It provides much biological detail, and also considers the role of spatial attention. Their study bases the switching on neuronal adaptation, but does not see a normative role for multistability as such. Instead, the authors relegate the functional relevance of adaptation to a role it plays during learning only.

'rational'). In contrast to our own approach, their model is both conceptual, with the variable to be inferred and sampled explicitly representing the orientation of the cube, and high-level, in the sense that no further claims are being made about the implementation of the algorithm in the brain, say in terms of neuronal mechanisms. The distinction between explicit and more implicit variables as in the DBM also carries over to the way the bistability problem itself is framed. Our model never uses an explicit description of a bimodal posterior over an underlying abstract, low-dimensional variable. Rather, the whole bistability problem is implicit in the very high-dimensional posterior formed by the hidden units, with a distribution that is inferred to be bimodal in that space due to the representations learned in training. This is also more than just a conceptual or philosophical difference. The two resulting process models consist of algorithms that sample either in a very low-dimensional or a very high-dimensional space. Presumably, either models could make quite different predictions if compared in detail. On the other hand, it could also be the case that a low-dimensional model such as Sundareswara & Schrater's just fits the role of a more abstract, high-level equivalent of a more detailed (if still highly idealised) implementation such as ours, and that overall predictions could be similar. It would be useful to tease apart these issues further in future work.

In summary, we suggest that our model differs significantly from that of Sundareswara & Schrater in that the former can be characterised as being internal, instrumental, and mid-level, whereas the latter can be described as internal, conceptual, and high-level.

A 'rational process model' of binocular rivalry

It was Gershman et al. (2009b) who emphasised in their computational study of binocular rivalry that MCMC could have the key properties to account for the aspects of bistability, while being a more versatile, "rational" method for probabilistic inference in general. It should be noted that their model does not involve any adaptation, and might thus fail to fully account for the dynamics of bistability in binocular rivalry (to be discussed below, Section 5.5.3). They argue that this lack of appeal to ad-hoc neuronal mechanisms to explain bistability is a strength of their approach. However, once one makes the general notion of MCMC in the brain more concrete, it could simply be the case that neuronal adaptation could itself be a (possibly heuristic) mechanism of how MCMC is actually implemented in neuronal circuits (a point they make as well, Gershman et al., 2012). This is the view we substantiated with our model here.

Gershman et al. characterise their approach as 'rational process model' (Sanborn

et al., 2010), suggesting an interpretation of their full Bayesian model as external, but of the approximate sampling-based inference as internal. The involved variables explicitly describe the perceptual inference problem thought to underlie binocular rivarly, namely deciding between two alternative 'causes' that generated each pixel in an observed image, and the model has a topology that matches the stimulus (e.g. a ring to observe travelling waves in the percept). Their approach is not concerned with the neuronal implementation in the brain, in contrast to ours, thus rendering it purely high-level. Still, at first glance it might appear that their model and ours actually have a lot in common: both ultimately entail sampling in Markov random fields (the DBM is an instance of the latter). Nevertheless, the conceptual distinction w.r.t. what the variables represent (local binary causes of images in their case, neuronal representations in ours) remains. In particular, while both models happen to use binary variables, it is only in their case that these variables semantically map directly unto the perceptual alternatives in rivalry. In our model, what matters is that the posterior across the whole set of hierarchical distributed representation is bimodal. In principle, we would expect similar results even if we were to use non-binary units (e.g. rectified linear units, Nair & Hinton, 2010). In conclusion, their approach has both external and internal components, is conceptual, and high-level.

As a further note, in a very recent follow-up paper (Gershman et al., 2012) to their earlier study (Gershman et al., 2009b), Gershman et al.'s approach appears to shift towards using (or interpreting) Bayesian models as *internal* models. Again they model binocular rivalry as MCMC inference in a Markov random field, but now they use a different architecture for the latter (unlike earlier they also employ Gibbs sampling now, like our model). Importantly, the graphical model is stated to represent *the brain's* assumptions about how the sensory input was generated. In particular, the generative model includes an outlier process that allows inference to ignore sensory data coming from one of the eyes if necessary. Thus, the graphical model itself, and not just the approximate inference algorithm, aims to capture representations and processes internal to the brain. Lastly, the explicit mechanism to ignore one of the eyes if necessary is an interesting aspect, which is not explored in our rudimentary model of binocular rivalry so far, among several others (see Section 5.5.6 on respective future work).

Bistability from sampling in neuronal circuits

Another recent example of a sampling-based account of bistability is that of Moreno-Bote et al. (2011). Their modelling and experimental study is concerned with a stimulus

consisting of two superimposed moving gratings, where the apparent depth ordering of the two patterns is ambiguous and induces bistable perception. The ordering is influenced by cues such as the speed and wavelength of the individual gratings, allowing for a more detailed analysis of how these cues interact with bistability.

Their work can be separated into two parts, with the second being the most relevant for us here. Briefly, in the first part, they make an argument in favour of the notion that the brain might implement sampling-based inference. They show how aspects of their experimental results, regarding how the depth cues influence the fraction of overall time spent in either of the percepts, match a 'multiplicative rule', which they derive from theoretical considerations of the underlying probabilistic inference.[11] Their analysis is here rather high-level and does not consider the algorithm by which samples are drawn at all, including its possible temporal dynamics. In the second part, Moreno-Bote et al. provide a concrete neural implementation that has the right properties for their general analysis made earlier to apply. They map inference over a binary variable (the depth ordering to be inferred in the task) directly onto two competing populations of neurons, modelled in terms of their global firing-rates, and include both noise and an adaptation process. Thus, unlike the approaches discussed so far, this model is low-level (as well as internal). At the same time, it is also conceptual unlike the DBM, as the neuronal activity explicitly corresponds to the perceptual variable in question.

Moreno-Bote et al.'s approach might also exemplify potential problems with attempting to find the neuronal substrate of Bayesian inference by mapping abstract conceptual entities, such as the depth ordering variable, explicitly onto neuronal representations. First, their implementation is rather specific to inference over a single binary variable, and it is not clear how it generalises to other vision problems. Second, as their network corresponds to two idealised interacting populations of neurons (the model could equally well describe two individual neurons), it likely skims over a lot of the essential details of how this form of bistable perception is realised in the cortex. Our model still leaves much to be desired in both points as well. However, we would argue that what speaks in its favour are its relation to the more general framework of MCMC-based inference, and the fact that it uses hierarchical distributed neuronal representations likely to be important for explaining the cortical basis of bistable perception (e.g. Tong et al., 2006)

Third, in the case of their model, the sampling hypothesis loses its normative moti-

[11] The reader interested in their study is encouraged to consider its supplementary material, as the main paper might be somewhat implicit w.r.t. the generality of their results, for example concerning the involved 'multiplicative rule'.

vation. As the authors write in the introduction to their paper, sampling is useful where complex probabilistic inference is otherwise intractable. However, once the perceptual problem is assumed to be solved up to a stage where its description is reduced to a single binary variable, which is then explicitly represented in two neuronal populations, there is little benefit in using sampling to approximate the posterior. Rather, the posterior and subsequent computations based on it could be represented exactly. For comparison, in our model sampling is useful because it happens not w.r.t. a low-dimensional abstract variable capturing explicitly the specific perceptual problem, but over a high-dimensional latent representation of images.

A hierarchical model of binocular rivalry

Finally, the model of binocular rivalry of Dayan (1998) perhaps comes closest in terms of the overall approach to our own. There, binocular rivalry results from inference in a generative hierarchical model with a neural network interpretation. The rivalry itself stems from a competition between the alternative generative explanations of the sensory input (corresponding to the two individual images). That only one interpretation becomes dominant is ultimately however a product of approximate mean-field inference, which, when the approximate posterior is constrained to be unimodal, fits only one mode of the true posterior and ignores the other.[12] To then obtain perceptual switching, an additional fatigue process is implemented (which is not further motivated normatively). Even then, inference still remains deterministic however and thus lacks the stochasticity observed in perceptual bistability. Dayan thus pointed out the possible relevance of the inherently stochastic MCMC methods such as Gibbs sampling, but also noted that "it is not obvious how to incorporate the equivalent of fatigue in a computationally reasonable way". In a sense, this is what our work addresses, although the adaptation

[12] Similarly, the predictive coding model in Friston's free-energy framework (Friston, 2009) has been related to binocular rivalry (Hohwy et al., 2008), and it also employs a unimodal mean-field approximation to the true posterior. Hohwy et al. (2008) also offer the explanation that perceptual switching could fall out in a Bayesian framework on the basis of a prior expectation for changes in the environment (suggested earlier by Bialek & DeWeese, 1995). They do not provide however a concrete computational model. We note that their account of binocular rivalry appears to us to be not completely clear about where the selection of a single hypothesis stems from. In that paper, they appear to argue that the selection directly follows from exclusivity of the alternative causes: "We [primates/humans] have therefore learned that the explanation for binocular visual input is unitary (i.e., has just one cause)." However, this on its own should yield a suppression of one cause in the posterior (via explaining away or negative correlations) only once additional evidence favours one of them over the other. With evidence and prior for both being equal, the posterior should still be bimodal and give equal weight to both. On the other hand, if Friston's actual computational model were to be used (Friston, 2009), the unimodal mean-field approximation should achieve the desired effect.

mechanism still remains somewhat of a heuristic (an aspect of the underlying rates-FPCD algorithm).

Dayan's approach is clearly framed as internal model, describing how 'cortical explanations' compete and hence result in bistability. At a similar or possibly somewhat lower level of description compared to the DBM, the goal is to map the graphical structure of the model onto neuronal units and their connections, and inference onto neuronal dynamics. While the units in the model are manually designed (via their receptive fields) to signal the presence of interpretable features (horizontal or vertical bars), they are understood to be representations that would normally be learned from images, and they carry only weak conceptual meaning. Ultimately, for Dayan's argument it does not matter what these units represent as long as they compete for 'explaining' the input. Thus, this model is internal, mid/low-level, and likely instrumental.

There is one more issue relating to the conceptual status of a Bayesian model that we can address with the help of the example of Dayan's model. His full probabilistic model is defined over n binary variables or units that together represent an image. However, a full representation of the corresponding posterior distribution would require 2^n values. Hence, to map these variables to actual *neuronal* units requires an approximate representation. In our own work, the corresponding mapping was realised by assuming that the neuronal network represents a single sample, a single point in the high-dimensional space. In Dayan's model, the mean-field approach leads to the posterior factorising into n separate terms. The thing to note is that his full generative model is not actually meant to be implemented in the brain, rather the approximate one is; nor does the former describe the external world or an ideal observer, which is how 'rational' approaches are often motivated. Arguably, in such a case, finding that a biological system can be described as approximately implementing inference in such a full Bayesian model can provide a useful theoretical perspective. However, without the relation to an ideal observer or truthful world model, this comparison is not necessarily unique or privileged.

5.5.2 Relating probabilistic representations to visual experience

In this section, we briefly summarise the overall issue of how approximate probabilistic inference in the brain relates to both cortical processing and the perceptual experience that people report, addressing the corresponding aspects of the aforementioned models and of other relevant approaches. For bistable perception, framed as inference with a

bimodal posterior, the fact that only one of the alternatives ever seems present in awareness concurrently is usually taken as evidence that, at least on this high level, people do *not* represent full multi-modal posteriors at the same time (Hoyer & Hyvärinen, 2003; Gershman et al., 2009b; Sundareswara & Schrater, 2008). This thus has implications for the inference algorithms used by the brain, and can be addressed in several ways.

The first possibility is to identify the current state of the brain with a single point in the space of the posterior, thus explaining why only one possibility is experienced at a time. This is the view taken in particular in our approach and Gershman et al.'s, which explain bistable perception with MCMC sampling, where the current percept as reported by the subject can naturally be related to the current sample produced by the Markov chain. However, for MCMC to make sense normatively in the context of probabilistic computations, the generated samples would need to be integrated at *some* point, maybe in prefrontal cortex (Gershman et al., 2012). For bistable perception, this appears to be in line with intuition, as the subject's knowledge about the set of possible explanations of the sensory input seems to be not instantiated in momentary visual experience but rather rely on working memory. It is however unclear whether this representation of uncertainty at a rather high, cognitive level would be in line with the various Bayesian accounts that see uncertainty being dealt with more generally throughout the brain, including in perception at various stages, motor control, etc. (Knill & Pouget, 2004).

Second, the posterior in the brain might be approximated to be unimodal, as is the case in the variational mean-field approximations in Friston's predictive coding model (Friston, 2009, to be discussed in Section 7.1) as well as Dayan's model of binocular rivalry considered above (the approximation is Gaussian in the former case, factorial over binary variables in the latter). On the other hand, such accounts then might need to pose additional mechanisms to explain the dynamics and stochasticity in bistable perception (Dayan, 1998; Hohwy et al., 2008).

Another option for a principled approximate inference algorithm that might combine a single sample representation and stochasticity with a basic measure of uncertainty (not requiring accumulation of samples) is a particle filter. In cognitive science, the latter has been proposed to describe human inferences in 'rational process models' (Sanborn et al., 2010). Evidence suggests that human inference is often well-described as using just a single particle (op. cit.). Again this would correspond to a single point in the space of the posterior, however with an associated weight that, depending on how the particle filter is defined (Daw & Courville, 2008), can reflect a measure of uncertainty or agreement of the currently entertained hypothesis with the incoming

sensory data. Similarly, in Yu & Dayan's model of acetylcholine (Yu & Dayan, 2002, 2005; Section 4.1.1 and Section 4.4.3), only a single high-level hypothesis is entertained at any point in time, with an additional measure of associated uncertainty coded for by the level of acetylcholine.[13]

Finally, the brain might also employ disparate algorithms or represent uncertainty differently in different contexts or anatomical systems (Vilares & Kording, 2011). In bistable perception, experimental evidence shows that neuronal activity correlates with the current percept most selectively in higher cortical areas, whereas modulation in lower areas is weaker (Tong et al., 2006). Based on evidence like this, Lee & Mumford (2003) suggested that the brain might use a particle filter based representation, but one where the activity in higher areas only represents one or a few samples from the distribution over variables represented there, while lower areas represent a larger set of concurrent samples, but these do not enter the subject's awareness, maybe due to interactions with prefrontal cortex. Similarly, in the model of Sundareswara & Schrater (2008), a whole set of samples from the bimodal posterior over Necker cube interpretations is represented in the brain, but only one of them is selected to be brought into "awareness and memory" (as mentioned earlier, this selection mechanism might be somewhat ad-hoc). Still, a gradual effect of bistable perception on the cortical hierarchy could also be explained with single-sample models such as ours. In that case, the reason that the activity of lower areas correlates less with the current percept could be either, that the information or variables represented in those areas are not actually those subject to ambiguity (the latter being depth or object identity in the case of the Necker cube); or, because there is a gradient along the hierarchy in terms of how strongly ambiguity or conflicts in the input are resolved by overriding them with internal explanations. The latter might be how our model could be interpreted.

Ultimately, these different approaches–particle-filters with single particles, MCMC, unimodal variational approximations–might have significant commonalities. Possibly, they all can be seen as different forms of idealisations that provide distinct perspectives on what is really taking place in the brain. In particular, they might correspond to idealisations along different dimensions of our scheme for categorising Bayesian models. For example, a particle filter could be a conceptual, high-level description of inference in terms of a low-dimensional variable explicitly representing a relevant property of

[13]The single-particle filter model of Daw & Courville (2008) is actually based on the model of Yu & Dayan (2005). It should be noted that, much like MCMC that does not accumulate samples, a particle filer with a single particle is very far from ideal normatively, but it could still provide a useful theoretical framework to characterise human observers.

the world. To match this high-level description to distributed neuronal representations in the cortex, a more instrumental type of model might be needed that might run a somewhat different algorithm, such as MCMC, perhaps with an additional measure of uncertainty associated with the current hypothesis (with this measure then represented, for example, with neuromodulators, Yu & Dayan, 2002). Exploring these connections further in future work could be highly fruitful.

5.5.3 The need for both noise and adaptation in bistability

The aforementioned neural network model by Moreno-Bote et al. (2011), involving competition between neuronal populations modelled in terms of their global firing-rate dynamics, is an example of a type of model commonly applied to bistable perception from a more mechanistic, non-probabilistic perspective (e.g. Wilson, 2007). Of high relevance to our work presented here is the modelling study by Shpiro et al. (2009): they argue that in order to account for the experimentally observed statistics of bistability, *both* noise and adaptation are necessary. Recent psychophysical work by Kang & Blake (2010) further corroborates this claim.

Our approach involves both noise inherent to MCMC sampling as well as neural adaptation to enhance the exploration of the posterior. While both noise and adaptation might play a role in the brain, an interesting question to ask is whether both are strictly necessary for bistability in our model as such. As we have already mentioned earlier, for the specific model instances used (with parameters learned with basic training methods), neuronal adaptation was necessary to escape the currently inferred mode and evoke bistability. In a few preliminary experiments, we also explored the importance of noise in the model. The hidden units were no longer sampled as normal, but instead the real-valued unit activations were propagated through the network (which without adaptation would correspond to mean-field inference, Section 2.2.2), thus rendering inference deterministic.

For the binocular rivalry setup, we interestingly found that without sampling noise, adaptation was indeed insufficient to evoke switching of the percept; the decoded hidden states would converge to one of the alternative images and remain there. Adaptation was still affecting the hidden units however, as subsequently reinitialising inference over the same input (keeping the adapted neuronal parameters from before) immediately lead to convergence to the alternate percept. Adaptation thus was not strong enough to evoke switching between the modes in the posterior–that is, for the adaptation parameter

values chosen in this study. The latter were originally determined such that the percept over non-conflicting images (as used in training) was stable, whereas bistability was evoked in the rivalry condition. Now, without noise, we found that stronger adaptation indeed lead to switching in rivalry, but then the normal percept was rendered unstable as well.

While this is only a preliminary exploration in the model, it possibly reveals a normative explanation for what Shpiro et al. (2009) found in their model on phenomenological grounds, namely that bistability should operate with a finely tuned balance between noise and adaptation. In our model, both noise and adaptation can be useful for exploration of the posterior. However, the requirement of stability for unambiguous or non-conflicting input puts limits on the extend of neuronal adaptation in particular, as unlike sampling noise, adaptation can strongly reshape the energy landscape and thus the whole posterior as such.

On the other hand, we found a different outcome for the Necker cube simulation. Here adaptation was sufficient to evoke bistability even in the absence of noise, at least for the parameter setting used in the remainder of this study. It should be noted that the psychophysics study by Kang & Blake (2010) that argued for the necessity of both noise and adaptation was concerned with specifically binocular rivalry, not the Necker cube. In any case however, more work would be necessary to see in how far our results are not only coincidental to the details of the model setups used. The aforementioned insights about the limits of adaptation in the brain might be useful regardless.

5.5.4 Synthesis in bistable perception

Similar to other approaches (Dayan, 1998; Gershman et al., 2012), we modelled bistable perception in the context of a generative model, where bistability arises from the system exploring different hypotheses that are compatible with the sensory input. This can be described as analysis by synthesis in the sense that higher cortical areas are thought to actively construct a top-down explanation of sensory input that is evaluated by some analysis procedure (Dayan, 1998). But what exactly is being synthesised in perceptual bistability?

Binocular rivalry could be conceptualised merely as a selection process on the level of inputs rather than that of explanations, so the question of synthesis is more of interest in the case of the Necker cube. Here, a 2D pattern[14] of lines is translated by the brain

[14]The author notes that it appears to him there is indeed a third perceptual interpretation of the Necker cube as such a 2D pattern.

into a 3D wire-frame cube in the act of interpretation. In a generative model of visual scenes, this could be captured by the filling-in of missing depth information. On the other hand, it could be argued that in the case of a drawing of the Necker cube on a flat piece of paper, depth information, from disparity or other cues, is not so much missing but rather in actual conflict to the cube interpretation found by the brain.

How is a paradoxically 'flat 3D cube' represented in the brain? In a hierarchical architecture consisting of specialised areas, this might be realised by having a high level area that codes for objects (e.g. area IT in the cortex) represent a 3D cube, whereas a low area that is primarily involved with depth cues as such represents a flat surface. An intermediate area might be biased towards a 3D interpretation via top-down input from the high area. Such a coexistence of partially contradicting information across the hierarchy is also present in our DBM model (see e.g. Figure 5.3), and will become a major point of interest in the chapter on attention (Chapter 6).

5.5.5 Preliminary experiments with depth information

The above argument helps somewhat to justify our simplified setup. Not only did we not model depth explicitly, but we also had the model trained on unambiguous opaque cubes, which later on could be inferred as the interpretations of the Necker cube. Clearly, opaque cubes are not quite the same as wire-frame cubes with depth information, and the bistability in the model essentially arises because alternative edges of the cube are ignored to arrive at the opaque interpretation. However, in light of the above argument, opaque cubes could be seen as an idealisation of reality under the assumption that the 3D interpretation of the Necker cube image similarly involves ignoring or overriding actual depth information, rather than just filling in thereof.

Nevertheless, treating depth explicitly in the model would be more appropriate and could be a line of future work. We performed initial experiments in that direction. Using real valued visible units, we trained a DBM with two images per stimulus modelling 3D wire-frame cubes: one image contained a binary wire-frame cube representing 2D information, and one consisted of a gray-scale image representing the depth of the pixels in the first image (Figure 5.10a). We then tested the model with input where the depth of the cube had been set to be flat, corresponding to a 2D drawing of a Necker cube (Figure 5.10b). We can report that in a few pilot experiments, the DBM not only inferred depth in higher layers (in contrast to the actual depth input signalling a flat cube), but also exhibited bistable perception when neuronal adaptation was employed.

(a) Cube with depth. (b) Flat (Necker) Cube.

Figure 5.10: Example images from preliminary experiments using explicit depth information. Images were split in two halves, where the left always contained binary wire-frame cubes corresponding to non-depth, 2D information, and the right contained a real valued version of the same cube, representing the absolute depth of the corresponding pixels. The model was trained on 3D cubes *(a)*, and then tested on a flat ones *(b)*, i.e. Necker cubes. In response, the model was found to infer depth in higher layers regardless, and to exhibit bistable perception (not shown).

This setup seems much closer to reality, and could be further extended in future work by learning about depth using actual stereo images rather than explicit depth information, as has been done in very recent Boltzmann machine models (Memisevic & Conrad, 2011).

5.5.6 Future work

Our work on bistable perception was part of the overarching study of perceptual phenomena in the generative DBM model, as presented in this thesis. It thus might lack the depth of other models dedicated to this subject alone, and thus there is much opportunity for future work. Part of the focus should lie on more extensive match to experimental data. One might be cautious about the utility of precise quantitative reproduction, considering that even if the DBM does share computational principles with the cortex, the actual implementation of course differs. Most useful would thus be further simulation experiments that elucidate on which experimentally found phenomena can qualitatively be explained on the basis of the underlying computational principles, and which cannot.

Examples of future lines of work already mentioned are: the inclusion of depth information or even stereopsis to model depth perception for the Necker cube (Section 5.5.4), which is of course relevant for binocular rivalry as well; a detailed analysis of the role of stochasticity and adaptation in the model (as in the work of Shpiro et al., 2009, Section 5.5.3, or in that of Kim et al., 2006); and, an exploration of whether sampling in the distributed hidden space of the DBM can be related to sampling algorithms

formulated in terms of more abstract, conceptual high-level variables, such as in the model of Sundareswara & Schrater (2008) or in particle filters (Sanborn et al., 2010).

Other examples would be based on testing whether various additional phenomena can be reproduced in the model, in particular those considered by other models. For instance, these could the role of contrast in binocular rivalry (Dayan, 1998), 'travelling waves' in the percept (Gershman et al., 2009b), or the influence of scene context on the interpretation of the Necker cube (Sundareswara & Schrater, 2008). For the latter, the capability of the DBM to learn from the statistics of sensory data might be particularly relevant. In all cases, it would be particularly interesting if different models were found to make different predictions.

Finally, to emphasise the importance of neuronal adaptation, it could be very useful to reproduce in detail the psychophysics experiment of Kang & Blake (2010), on the basis of which the authors concluded that both noise and adaptation are necessary to account for the statistics of binocular rivalry. The key element of this experiment was to intersperse the presentation of conflicting input images with periods where only one of the input images was presented, thus allowing for a controlled prolongation of the duration of potential adaptation to that image. Reintroducing the second image then lead to an increased likelihood of switching to the corresponding other percept. Carefully controlling the experimental conditions, the authors argued that this effect is difficult to account for with models that rely solely on noise. Again, we can report from preliminary experiments that these results seem to be reproducible in our model at least in principle. It would be straight-forward to quantify this in future work.

5.5.7 Conclusion

Our work established a possible connection between mechanistic accounts of perceptual bistability in terms of neuronal adaptation and normative accounts in terms of sampling-based inference methods. It thus contributes to the ongoing trend to explain cognitive processing in terms of approximate probabilistic inference, and demonstrates further the relevance of hierarchical generative neural networks such as the DBM. There is much room for future work, and we have already performed preliminary experiments on including depth information for the Necker cube and on prolonging periods of adaptation in binocular rivalry (Kang & Blake, 2010).

Chapter 6

Generative feedback processing for object-based attention

The idea behind the brain implementing a generative model is that it could learn from, and make inferences about, sensory data on the basis of internally synthesised explanations, from which the sensory input can be generated or predicted. In the case of vision, creating a generative model of images of even simple objects–even binary digits–is not an easy task. The brain however has to deal not only with the inherent richness of visual data, but also with the added complexity of visual scenes being composed of many objects. And, moreover, it needs to meaningfully combine visual information with data from other sensory modalities, and then form internal representations on the basis of which planning, memory, and action can operate.

Attention can be framed as a possible solution to this challenge. After all, our brains do not solve all perceptual problems in our sensory environment at the same time. Even when our bodies are static and our eyes fixate to one location, we can still shift an internal focus of attention covertly to give chosen aspects of the sensory environment privileged cognitive access. According to some theories of attentional processing in the brain (e.g. Rensink, 2000), paying attention to a visual object in that manner corresponds to a specific form of representation in the cortical hierarchy, where higher areas represent primarily the object currently in the focus of attention. Recurrent processing and feedback could then serve to bias representations across the cortex towards the attended object, emphasising relevant and/or suppressing irrelevant information, and binding distributed representations into one coherent percept (e.g. Duncan et al., 1997).

In this chapter, we take some early steps towards modelling object-based attention

in the generative framework of the deep Boltzmann machine (DBM). There are two angles to this work. From a biological modelling point of view, the key idea is to model an aspect of attentional processing by making use of the capabilities of the model to synthesise internal representations. We show that by employing a model which has learned representations specific to single objects only, inference over images with multiple objects can effectively implement a form of attentional selection realised by recurrent processing. From a machine learning point of view, the issue to explore is how Deep Learning approaches can be related to attention, and what kind of inference mechanisms and representations are necessary.

To establish a connection to biology, we qualitatively elucidate on the following aspects of theories of attentional processing in the cortex using the DBM model: first, the notion of a fast feedforward (FF) sweep followed by subsequent recurrent processing, the latter being essential for perceiving objects when scenes are cluttered (Lamme & Roelfsema, 2000); second, that in directing attention to an individual object in a scene, an attractor state is assumed which binds together and emphasises aspects of that object represented throughout the cortical hierarchy, suppressing representations of competing objects (Serences & Yantis, 2006; Tsotsos et al., 2008); third, the hypothesis that scene representations in the cortex are inherently such that higher stages represent primarily one object at a time, unlike lower stages such as V1 where the whole image is encoded in terms of low-level features (Rensink, 2000).

Our main focus is the biological application, but on the technical side we show how deepness of the architecture and restricted receptive fields are important for realising the attentional state, making the DBM robust against noise not seen in training. Finally, we investigate additional suppressive attentional mechanisms to cope with problems beyond toy data, and argue that sparse representations could be critical to that end. In this context we explore a spatial attention mechanism in the form of an internal 'spotlight' as well.

These results, as published (Reichert et al., 2011a),[1] represent early work that demonstrates how attentional processing could be understood in the DBM framework, and what the implications are for neuronal representations and hierarchical inference in such models. In addition to the main results to be presented below, we also explored means of extending the model with the goal of addressing some key issues with the work so far. The latter are: finding a more principled model formulation, making the

[1] The simulation results are mostly as presented in that paper, but much of the description has been rewritten.

model work with more challenging data, and realising learning about individual objects from complex scenes. The attempts made have been inconclusive for now, but will be described in Section 6.5.3 as part of the discussion.

In the next section, we give a brief overview over aspects of object-based attention, such as the interplay of attention and object perception. We also discuss the Selective Tuning model of Tsotsos (2011b) and Bayesian approaches to attention. In Section 6.2, we describe the DBM model of object-based attention and the data sets and model setups used throughout this chapter. Section 6.3 covers our simulation experiments using the basic DBM model. In Section 6.4, the model is extended with additional suppressive mechanisms, also allowing for modelling spatial attention. We conclude in Section 6.5, discussing our results, issues such as the role of invariant representations and learning from motion, and preliminary and future work.

6.1 Attention in the cortical hierarchy

Attention (for reviews, see e.g. Ward, 2008; Carrasco, 2011; Tsotsos, 2011b) is a very complex phenomenon, or indeed a label for a group of phenomena (Driver et al., 2001; Tsotsos, 2011b), possibly with a somewhat unclear conceptual status (e.g. Anderson, 2011). Doing the subject justice with an extended review and discussion is not possible here, hence we merely clarify relevant terms and address several key aspects and key computational models to establish the necessary context for our own work.

A lot of work on attention has focused on the mechanisms by which the brain might decide where to direct, or orient, attention in an image, e.g. to coordinate eye movements. Such a decision could be based on saliency maps that are computed in a bottom-up, externally driven fashion from image features (such as contrast in various feature dimensions, Itti & Koch, 2000; Bruce & Tsotsos, 2009), highlighting salient image regions that then draw attention. Additionally, conspicuous image regions could be computed in a task dependent fashion, e.g. when searching for an object with certain properties among clutter. Here, however, we are not interested in the question of where attention should be directed, but rather the issue of what happens in the brain *during* the act of deploying attention as such.

Moreover, our subject of enquiry is *covert* attention (Carrasco, 2011), where attention is selectively directed to parts of the sensory input while sensory organs such as the eyes remain fixed. A shift of attention in such a manner is correlated with behavioural and neurological changes, and thus implies an alteration of the state of the brain that

is not evoked by variation in the external input. The synthesis of an explanatory internal state that can change dynamically even when input is fixed is a common theme throughout this thesis, and will be our entry point to model attentional processing in the generative DBM model. But what are the neurological nature and functional role of covert attention?

Visual attention is commonly thought to be a selective process that is a consequence of limits to the capacity of the brain to process visual information. The nature of this capacity limit is a subject of debate, as is the question of how early or late in the visual system a selection of represented information occurs. In terms of the neuronal correlates of attention, evidence shows that there is generally a gradient along the cortical hierarchy, where neuronal activity in higher areas is most strongly affected and altered depending on what in the visual field is currently being attended, whereas effects in early areas such as V1 are weaker but still clearly measurable (see reviews cited above, or e.g. Serences & Yantis, 2006; Desimone & Duncan, 1995).

Attentional processing can be divided furthermore according to what aspect of the visual input is being attended. Spatial attention is directed towards one or multiple locations in the visual field. It is often described as an internal 'spotlight'. Feature-based attention emphasises certain elementary aspects of objects in an image such as colour or orientation, and does so globally across the visual field. And finally, object-based attention is thought to relate to an individual object and its structure beyond what is accounted for by a spatial spotlight or individual feature dimensions. In this work, we establish a connection to object-based attention, as well to a lesser degree to spatial attention.

6.1.1 Examples of object-based attention

The notion that attention can be object-based, rather than spatial or feature-based, is put forward by several psychological theories and some computational models, and is supported by a variety of evidence. As is the case with attention in general, object-based attention is a complex phenomenon and its different aspects are not always made conceptually distinct, nor is the distinction between it and other forms of attention necessarily clear (e.g. Rensink, 2000; Scholl, 2001).

The following are examples of the kinds of evidence that have been framed as relating to object-based attention (for reviews, see Vecera et al., 2001; Olson, 2001; Scholl, 2001; all examples here are in particular discussed by Scholl). Attention is

thought to provide certain stimuli with preferential processing, and psychophysical experiments have shown that processing advantages due to attention can automatically spread within objects, even if the latter are defined by illusory contours or are partially occluded. Moreover, attention can select one of two stimuli even if they overlap completely spatially, including high-level and complex stimuli such as superimposed movies or images of houses and faces. In the latter case, attentional selection is also accompanied by selective activation of the corresponding higher cortical areas that preferentially process either category of objects. That objects form the perceptual units on which attentional selection operates is also indicated by evidence from patients who suffer from Balint's syndrome as caused by parietal lesions. Some display what is termed simultanagnosia: it appears as if they cannot perceive more than one object at a time. In consequence, what they can perceive is determined by what can be grouped into an object e.g. according to Gestalt principles (see also Vecera et al., 2001).

6.1.2 Interplay between attention and object perception

Object-based attention differs from spatial and feature-based attention in that it relates to high-level visual processing rather than low-level and elementary aspects of images. Object-based attention is thus inherently connected to the mechanisms by which the brain segregates or segments parts of visual scenes into coherent objects. Addressing it is thus further complicated by our lack of understanding of how vision is realised by the brain. Many experimental studies (Carrasco, 2011) and computational models (e.g. Reynolds & Heeger, 2009) of attention focus on low-level effects in lower and intermediate cortical areas, such as contrast enhancement. Considering the functional importance that attention could have for higher level processing, and correspondingly the more extensive effects of attention in higher areas, these studies might be limited in how far they can elucidate the nature of attentional processing.

A first key issue is then whether it is object segmentation that influences attention, by determining what perceptual constructs can be attended, or whether attentional mechanisms themselves play a role in what is perceptually grouped as objects in the first place. This issue is discussed in detail by Driver et al. (2001)[2] and relates to the question of early and late attentional selection, and to the traditional dichotomy of preattentive

[2]The authors also suggest that the term 'object-based attention' suffers from the vagueness of what constitutes an 'object'. They argue that the relevant research would be characterised better as being concerned with precisely the interaction of segmentation and attentional processes.

and attentive vision. As Driver et al. argue, there seems to be at least some evidence for both, and the aforementioned dichotomies might be too simplistic.

Some psychological theories thus pose intermediate or hybrid accounts of the involved processes. For example, bottom-up, parallel sensory processing could lead to the formation of 'proto-objects', volatile representational constructs that serve as the units of attentional selection, based on some initial segmentation and grouping mechanisms (Rensink, 2000; Scholl, 2001; Driver et al., 2001; Collerton et al., 2005). It is then the attentional selection that promotes a proto-object into a full-blown object representation, which might consist of a more detailed, coherent and lasting percept that enters awareness and is available for memory and decision-making. For example, Rensink (2000) argues that at this higher level of processing, which maps to higher cortical areas, only *one* object is ever represented at a time. The phenomenological impression of perceiving a visual scene as a whole would be an illusion that is possible due to information being made available dynamically whenever it is requested, via shifts of attention. The connection between selective high-level representation and mid-level proto-objects is to be established by recurrent interactions, realising some form of coherence. Normatively, representation of one object at a time is suggested as a solution to the complexity problem inherent to visual scenes.[3]

Attentional processing has been described as competition between distributed object representations in the brain (Desimone & Duncan, 1995; Duncan et al., 1997; Vecera, 2000). In so far as attentional processes interact with object segmentation, it should be clarified that conceptually, there might be two different forms of competition and selection involved here: one could be described as selecting between alternative and mutually exclusive *explanations* of a given image content (e.g. deciding that a line should be part of one figure but not another, or selecting one organisation into figure and ground over another in an ambiguous image); the other selects between separate entities that are all present in an image as such (spatially separated or not as in the case of superimposed images), maybe more along the classical account of attention as selecting between, or *filtering* of, stimuli. Though conceptually different, several descriptive accounts and computational models see both processes, and with them more generally object recognition, segmentation, and attention, inherently intertwined, possibly because they are based on

[3]The notion of perceiving only one object at a time needs clarification. After all, people (who do not suffer from Balint's syndrome) can compare multiple simultaneously presented objects perceptually as well as spread attention across them for example in the multiple object tracking paradigm (Scholl, 2001). According to Rensink, the higher-level entity subject to capacity limits more generally constitutes a 'nexus' that can link even multiple proto-objects or represent part-whole relationships, but still somehow is a unitary coherent construct.

the same neuronal mechanisms (Bartels, 2009; Lamme & Roelfsema, 2000; Qiu et al., 2007; Tsotsos et al., 2008; Tsotsos, 2011a; Spratling & Johnson, 2004; Scholl, 2001).

6.1.3 The Selective Tuning model

An illustrative example of a relevant computational approach is the Selective Tuning model of Tsotsos (1990, 2011b), which justifies and derives attentional mechanisms from computational and architectural constraints of the visual system. The model is intended to not require concrete assumptions about the nature of cortical representations but only about global architectural principles, such as a pyramidal hierarchical organisation and increasing receptive field sizes. The key idea of how attention is realised is the following: after an initial feedforward activation of the hierarchy, top-level neurons with large receptive fields compete for representation of objects in the image. Starting from the winning neuron or group of neurons there, the competition is then iteratively propagated down the hierarchy. Essentially, each neuron selects through recurrent processing the most strongly activated neurons among its input in the lower level, suppressing the others. The result is that neural pathways and representation are tuned according to what is currently attended and supported both by bottom-up sensory information and selective top-down feedback.

With feedback assumed to be sufficiently precise, there are several functional benefits of this selective tuning: the suppression of irrelevant image information within the receptive fields of neurons, which is important when scenes are cluttered; the binding of distributed representations and grouping of represented image content; and, importantly, the precise localisation of the attended object as representations in lower levels are selected, where topographically organised neurons have small receptive fields and thus carry explicitly detailed spatial information. Hence, according to Tsotsos and the Selective Tuning model, attention, binding, and recognition are closely related (Tsotsos et al., 2008; Tsotsos, 2011a).

6.1.4 Dynamics of attentional object perception

According to Tsotsos, an initial feedforward pass is followed by recurrent attentional processing, possibly repeated across several iterations as demanded by the behavioural task (Tsotsos et al., 2008; Tsotsos, 2011a). Similar arguments have been put forward by others. For example, based on various neurological findings, Lamme & Roelfsema (2000) also argue that an initial ‘feedforward sweep’ is useful for fast recognition, but

that additional recurrent processing is necessary for attentional grouping and segregation of image content into coherent representations. They also posit that such recurrent processing might be necessary for yielding visual awareness.

The notion that a global attentional state is somehow realised by spatio-temporal coherence across distributed brain systems is also a common one (e.g. Duncan et al., 1997; Rensink, 2000; Serences & Yantis, 2006). Serences & Yantis (2006) for example refer to the coordinated activity across the cortical hierarchy as a 'coherence field' (in the spirit of Rensink, see above). A shift of attention would correspond to the brain state switching between attractor states, and such shifts could be initiated with transient control signals originating from parietal or prefrontal cortical regions. How such coherence is established is less clear, and possible mechanisms could rely on synchronising oscillatory firing, gating, or selective enhancement and suppression of neuronal activity.

6.1.5 Bayesian attention

Attention has also been examined in the context of probabilistic inference. Framing attentional effects in Bayesian terms as resulting from a top-down prior comes naturally where attention acts as a top-down bias on alternative image explanations, or when it can be seen as resulting from prior expectations, e.g. when spatial attention is directed to an image location after a cue informed the observer that a relevant stimulus is likely to appear there (Posner, 1980). In such cases on the other hand the distinction between attention and *expectation* as such might not be that clear (Summerfield & Egner, 2009). Moreover, a resource limitation inherent to many attentional tasks (e.g. visual search) seems not to automatically fall out of a Bayesian formulation without further assumptions about the approximations in the model and inference mechanisms the brain uses. Thus, in such cases the intuition of 'attention as a prior' might be captured by models where attentional mechanisms have mathematical equivalence to Bayesian priors, but one should keep in mind that they do not correspond to true statistical priors in such cases (see Whiteley, 2008; Chalk, 2012, for reviews and related discussion).

Examples where attention has been related to the specific model structure, inference schemes, or information representation in the brain are the following. Several authors posed that the resource limitation necessitating attentional processes consists in the broadness of receptive fields of sensory neurons (in turn constrained by their finite number), leading e.g. to crowding effects (Dayan & Zemel, 1999; Dayan & Solomon,

2010; Yu et al., 2009). Given those constraints, subsequent Bayesian inference might then be described from an ideal observer point of view, or use further approximations for the sake of tractability when implemented in the brain (Yu et al., 2009). Similarly, the model of Chalk (2012) assumes broad receptive fields and then sees attention as reward-driven optimisation of the resulting neuronal representation.

According to the model of Whiteley (2008), the brain approximates the true posterior over the variables underlying visual scenes to make inference tractable, in a way that corresponds to variational inference in machine learning (Section 2.2). The posterior is assumed to be factorial such that it might match the specialisation of cortical areas. Attention moreover is then thought to be realised by the brain as *refining* the approximate posterior with an additional 'attentional hypothesis', which mathematically has the form of another multiplicative factor in the posterior. As the approximate posterior is optimised to be close to the true one, this attentional hypothesis has the effect of biasing the approximation according to what matters in a given behavioural context. It constitutes a flexible mechanism that could fulfil various computational roles associated with attention as is demanded by a task.

Tractability of probabilistic inference appears also to be key for Yu and Dayan's model of the role of neuromodulators in attention (Yu & Dayan, 2005, related to their earlier work on acetylcholine, Yu & Dayan, 2002, as discussed in Section 4.4.3 in the chapter on hallucinations). They study an extension of the classic Posner task (Posner, 1980). Subjects need to infer which out of multiple cues needs to be attended as it is currently relevant for predicting the location where a target stimulus will appear (they also need to infer *how* relevant that cue is). This is described as probabilistic inference over, in particular, the identity of the relevant cue. For reasons of tractability, they assume that the brain only maintains a *single* hypothesis about the identity, rather than the full posterior (plus measures of associated uncertainty signalled by neuromodulators). Presumably, this restriction to tracking only a single hypothesis is why there is an element of attentional selection in the first place.

Finally, Chikkerur et al. (2010) pose that attention can be explained on the basis of the brain using a probabilistic model that is inherently framed as being concerned with only one object at a time (and additionally, it uses a factorisation according to 'what' and 'where' information in ventral and dorsal streams, respectively). They argue that the benefit of this simplified model is perhaps not (just) addressing the computational complexity of inference, but rather to make it possible to learn the model from limited sensory data in the first place. In their study, effects of attention are then described as

following from these assumptions built into the internal model of the brain. This view is closest to the one we adopt in this work here, as will be explicated in the following.

6.2 Object-based attention in a DBM

In this work, we explore the possibility of utilising attentional processing during perceptual inference in the DBM model. The motivation is threefold.

First, on the premise that generative models could relate to learning and inference in the cortex, as promoted throughout this thesis, could attentional processing be the answer to the challenging complexity of sensory data? How can attention, in particular *covert* attention, be framed conceptually and implemented in generative models?

Second, if attentional processing involves binding and selection of distributed representations specific to an object across the cortex, could such selection be implemented by making use of the capability of a generative hierarchy to provide rich and detailed top-down predictions?

Third, can an attentional inference scheme in the DBM be related to the various aspects of attentional processing in the brain discussed in the last section, among them in particular: a fast feedforward sweep followed by attentional recurrent processing (Lamme & Roelfsema, 2000; Tsotsos et al., 2008; Tsotsos, 2011b); the emergence of the attentional state as some form of object-specific coherence across distributed representations, corresponding to an attractor state of the system (Duncan et al., 1997; Rensink, 2000; Serences & Yantis, 2006); and, a gradient of the impact of attention on hierarchical representation, where effects are strongest at the top, possibly because there only the attended object is represented at any point in time (Rensink, 2000).

Our results to be presented mostly constitute a proof of concept. Many open issues remain, including developing a more principled theoretical framework, making a closer connection to biology, and considering learning that utilises attentional processing. We will address some of these issues in the discussion (Section 6.5), and in particular report on some further work undertaken in that context (Section 6.5.3).

6.2.1 Approach

The key to our work here is the idea that the processing in the cortex is inherently concerned mainly with the currently attended object (Duncan et al., 1997; Rensink, 2000; Chikkerur et al., 2010). As Duncan et al. (1997) puts it:

> For the sensorimotor network as a whole, the tendency is to settle into a state in which different brain systems have converged to work on the same dominant object, analysing its multiple visual properties and implications for action. This is the state that, at the behavioural level, corresponds to 'focused attention' on the selected object. At the neural level, there should be widespread maintenance of the selected object's representation, accompanied by widespread suppression of response to ignored objects.

According to this view, attentional selection not only occurs late to allow for decision making and action but is part of perceptual inference itself. At the same time, the effect of attention in the sensory hierarchy is graded (e.g. Serences & Yantis, 2006), with a transition from a detailed low-level representation of the whole sensory input in lower stages to object specific representations in higher ones (Rensink, 2000).

One possible way to conceptualise this is to consider the cortex as implementing a generative model that is inherently formulated in terms of a single object (Chikkerur et al., 2010). How the rest of the sensory input should be treated, as well as where the gradual transition in the hierarchy stems from, is not obvious. For our exploratory study here, we used a DBM that had learned representations specific to individual objects (by training it on images with single objects). We then tested it on simple 'scenes' containing multiple objects or background clutter, essentially treating the other content as additional noise not seen in training. A gradual transition was an emergent effect, as will be explained below. How the rest of the scene could be treated more explicitly and be part in learning as well will be discussed in Section 6.5.

Note that the underlying assumption here is that attentional selection is *inherent* to recurrent inference in the visual system, automatically focusing the internal state towards one object in a scene. A signal that biases and controls which object should be attended in the first place might come from areas external to the (ventral) visual cortex, such as prefrontal cortex (Serences & Yantis, 2006), and is not part of the model, though we will study a means of controlling the attentional state via spatial attention in Section 6.4.4.

6.2.2 Model setup and data sets

The setup of the main model is again familiar from other parts of this thesis. In particular, it used restricted receptive fields in the weights, the size of which increased from lower to higher hidden layers. Also, sparse activity was encouraged simply by initialising the biases to a negative value at the beginning of training. In this chapter, we will examine the role of receptive fields more extensively. To this end, we also employed

for comparison a second version of the DBM that had full connectivity between the layers. The model with restricted receptive fields will be referred to as RRF-DBM, the one without simply as *plain* DBM. The importance of sparsity will be examined further in Section 6.4.

Two simple data sets of binary images were used to train DBMs on individual objects (Figure 6.1). The first contained toy geometric shapes (three kinds) at various positions in 20x20 pixel images (denoted *shapes*). The second was the MNIST data set of handwritten digits (*MNIST*, 28x28 pixels). For the attention experiments, more complex versions of the images were then used as input (based on images from the test set for *MNIST*), containing either multiple objects (*multi-shapes*) or objects with background clutter (*shapes+clutter*, *MNIST+clutter*),[4] as shown in the figure.

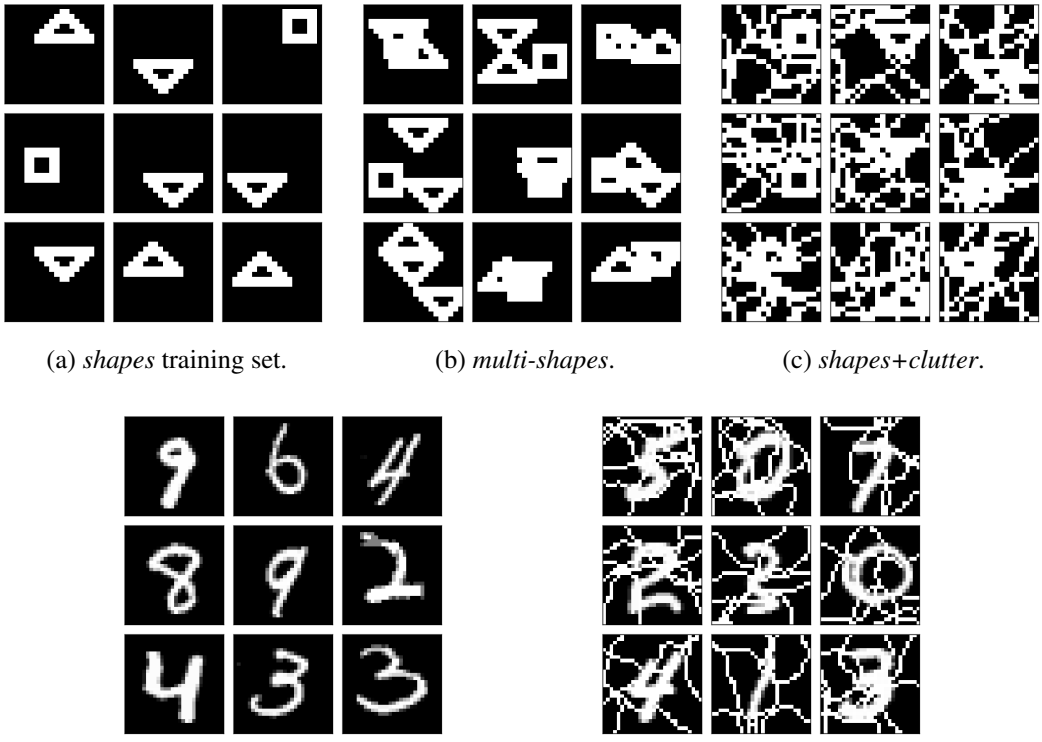
By applying our decoding procedure (Section 3.5), we could investigate whether the internal state of the model would relate to an individual object out of the scene in a fashion that could be related to attentional processing. For quantitative analysis, we examined how well the decoded image from the top hidden layer corresponded to an image containing only the attended object by computing the squared reconstruction error with regards to the latter. For the *multi-shapes* data set that contained multiple shapes, the attended object was simply defined as whichever shape was matched best by the internal representation. Lastly, we also attached a softmax label unit to the top layer (Section 3.5) to implement a classifier, seeing how well the attended object could be classified from the hidden representations.[5]

In detail, both plain DBM and RRF-DBM had three hidden layers. For the shapes images, these had 500/500/500 and $26\times26/26\times26/26\times26$ hidden units, respectively, where these numbers were chosen to balance the number of free parameters in the weights across both models. Receptive field sizes for the RRF-DBM were $7\times7/13\times13/26\times26$ (full connectivity for the plain DBM). For MNIST, the plain DBM had 500/500/2000 hidden units and the RRF-DBM $28\times28/28\times28/43\times43$, and for the latter receptive fields were of size $7\times7/14\times14/28\times28$.

Training used CD-1 for the shapes set and 5-step PCD for MNIST (30 epochs; see

[4]The background clutter consisted of randomly generated, irregular curves (six per image). Each curve was drawn as follows: draw an initial pixel and an angle ϕ, both uniformly distributed; draw two adjacent pixels in directions with angles $\phi_1 = \phi$ and $\phi_2 = \phi + \pi$; repeat: add to ϕ_1 and ϕ_2 increments δ_1 and δ_2, respectively, independently drawn from a uniform distribution over $[-0.5$ rad, 0.5 rad$]$; draw two new pixels adjacent to the respective last pixels, using the new directions.

[5]Note that this means that the top RBM in the DBM was not trained without supervision but instead used information about category labels, which could potentially affect the representations learned there. Relevant or not in the context of attentional processing, we did not find that it made a difference in practice in any case.



(a) *shapes* training set. (b) *multi-shapes*. (c) *shapes+clutter*.

(d) *MNIST* training set. (e) *MNIST+clutter*.

Figure 6.1: Example images used in the attention experiments. The *shapes* data set *(a)* contained simple toy shapes at various locations. Models were trained on it and then tested on images with additional clutter: *multi-shapes (b)* and the *shapes+clutter (c)* sets. Similarly, other models were trained on the *MNIST* digits (*d*) and then tested on digits with background clutter in the *MNIST+clutter* (*e*) set (based on digit images from the original test set).

Section 3.6.1 for further training parameters). All training data sets consisted of 60,000 images each (the more complex image data sets used for testing had 10,000 images each). At the beginning of training, the biases were initialised to -4 to encourage sparsity. To achieve sparsity, we had initially employed more sophisticated methods, such as regularising the biases during learning (Lee et al., 2008), but found that the resulting models were less capable of dealing with the complex versions of the images. We thus chose the simple bias initialisation instead. A more exhaustive analysis of the effect of these different methods on the model would be useful in the future.

6.2.3 Feedforward sweep and recurrent inference

To establish a connection to biological theories on the sequential nature of attentional inference in the cortex, we used both a feedforward sweep and recurrent processing during inference, as will be explained in the following.

In the cortex, an initial feedforward (FF) sweep of processing of a visual stimulus might sometimes be enough for quick recognition, but subsequent attentional processing realised through recurrent interactions is thought to be necessary when scenes are cluttered, both for recognition and to localise an object on a fine scale (Lamme & Roelfsema, 2000; Tsotsos, 2011b). In the DBM, perceptual inference is naturally recurrent (each intermediate hidden layer receives input from both adjacent neighbours), an advantage it has in this context over related models such as the Helmholtz machine (Dayan et al., 1995) or deep belief net (Hinton & Salakhutdinov, 2006, Section 2.1.4), where (approximate) inference is feedforward.

At the same time, the notion of an initial FF sweep is still sensible in the DBM framework. When they introduced the DBM, Salakhutdinov & Hinton (2009) utilised a single deterministic bottom-up pass to quickly initialise inference in the model, where each hidden layer received its input only from the layer beneath (the weights were doubled to compensate for the lack of top-down input). Intuitively, this allows the hidden units to be set in a way that is driven by the input, rather than starting normal (hence recurrent) inference with the network set to some state that is arbitrary with regards to the current stimulus.[6] Salakhutdinov & Larochelle (2010) extended this initialisation approach by using an additional set of dedicated recognition weights, the idea being to *learn* how to initialise the inference in the DBM to a good initial value. Even with this improved 'first guess', they showed that subsequent recurrent processing

[6] Such an initialisation would also be sensible in some form of online setting where the current model state might have been inferred over a preceding stimulus.

still was beneficial (there in the context of training the DBM parameters). Here, in our experiments, we used the simple bottom-up initialisation (doubling the weights) to make the connection to FF-sweep and subsequent recurrent processing in the cortex.

6.3 Experiments: attentional recurrent processing

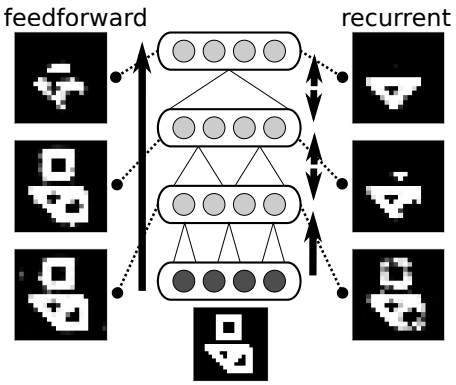
In this section we report our first set of results, with a focus on the roles of recurrent processing and the hierarchical organisation of the model. RRF-DBM and plain DBM models were trained on individual shapes or digits, and we then examined inference when the models were tested on complex versions of the images. Our findings will be illustrated with examples in the next section, and then backed up with a quantitative analysis in the section thereafter.

6.3.1 Emergence of an object-specific internal state

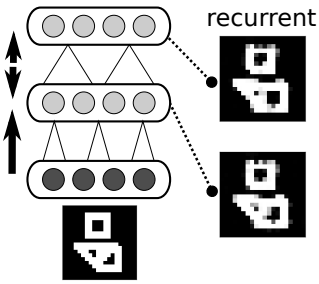
To begin with, Figure 6.2a shows example decoded hidden states for inference over a *multi-shapes* image in the RRF-DBM, comparing the internal representations after the initial FF-sweep and after subsequent recurrent processing (10 sampling cycles across the hierarchy). Decoded states are depicted for all three hidden layers. It becomes apparent that after the FF sweep, the hidden layer states were rather noisy, but the subsequent recurrent processing enabled the top layer to form a clearer representation of an individual shape, allowing both for a localisation of the object in image space and an improved classification (as shown later).

We indeed found a transition from representing most of the scene in lower layers to representing the individual object in the highest layer. Representations were biased towards the attended object even in lower layers, but this resulted from feedback from higher layers, as can be seen in the example by comparing the reconstructions of the first two hidden layers after the FF sweep and after recurrent processing. Only after the latter had taken place, involving feedback from the topmost layer, were the representations biased toward the individual shape. In fact, when we removed the topmost layer of the RRF-DBM, no object specific state was assumed (6.2b). This resulted in part because, due to the receptive field sizes, only the topmost layer had learned that training images only ever contained one shape.

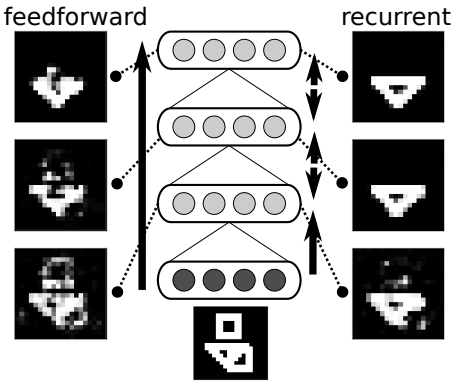
However, the deepness of the architecture played a role in itself as well: we found that even for the plain DBM, a model with two hidden layers instead of three performed



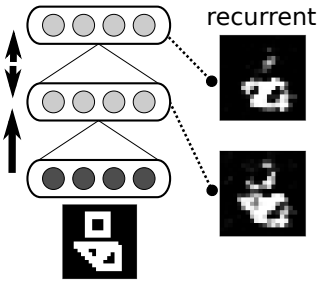
(a) RRF-DBM (3 layers).



(b) RRF-DBM (2 layers).



(c) Plain DBM (3 layers).



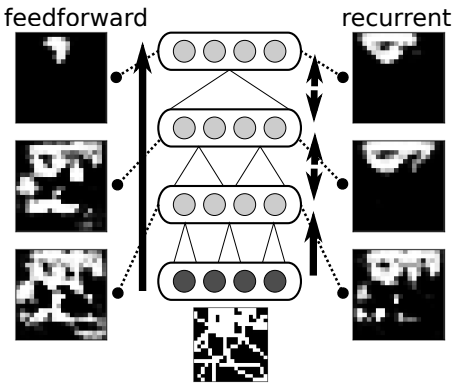
(d) Plain DBM (2 layers).

Figure 6.2: Example decoded internal states for all hidden layers of various versions of the DBM. The models had been trained on single shapes and were then tested on images containing multiple shapes to model attentional selection of individual objects. *(a)*: RRF-DBM (using restricted receptive fields) with all 3 hidden layers. Recurrent processing (right-hand side, after 10 cycles) improved attentional selection over an initial FF-sweep (left). *(b)*: with only two layers, the RRF-DBM failed to assume an object-specific internal state, even with recurrent processing (decoded images after the FF-sweep were here equivalent and are not shown). *(c+d)*: as (a) and (b) but with a plain DBM (using full connectivity between layers). Overall, the examples show that deepness and recurrency were important to realise the attentional state, the latter especially in the RRF-DBM: there, only the top layer had large enough receptive fields to learn representations specific to one object. Selectivity in lower layers was a result of feedback, and the resulting transition of selectivity in the hierarchy was more gradual in the RRF-DBM than in the plain DBM.

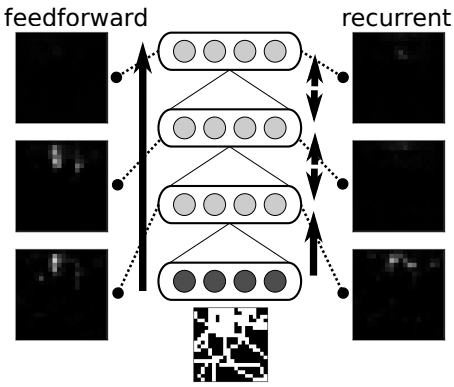
worse when it came to assuming an objected specific state (e.g. 43% vs. 22% classification error on *multi-shapes*; example case shown in Figure 6.2d). This was the case despite the fact that the hidden units in the plain DBM always saw the whole input image, and that in the three layer version, object specificity was actually more strongly expressed than in the RRF-DBM (Figure 6.2c), and even when the total number of hidden units in the two layer plain DBM was set to be the same as in the three layer version.

This finding is explained as follows. By virtue of how the model was trained, the object-specific representations correspond to ravines in the energy landscape and to (stochastic) attractor states of the system, at least if the model is sampling generatively or performing inference over actual training images. Conditioned on the more complex test images not seen in training, the nature of the inferred posterior is less clear. However, empirically it seems that hidden layers higher up in the hierarchy have a tendency to fall into states corresponding to what was seen in training, merely because they are further removed from the actual sensory data (independently of the receptive field structure), thus allowing the model to infer what it 'wants to see'. This of courses matches the findings in the other parts of the thesis, i.e. the representation of unambiguous interpretations of the Necker cube in higher layers (Chapter 5; e.g. Figure 5.3 on page 117), or even of hallucinations in the absence of input (Chapter 4; Figure 4.5 on page 75). Moreover, the transition to object-specificity is more gradual in the RRF-DBM, as there only higher layers have learned representation consisting of single objects, and selectivity in the lowest hidden layer stems mostly from feedback.

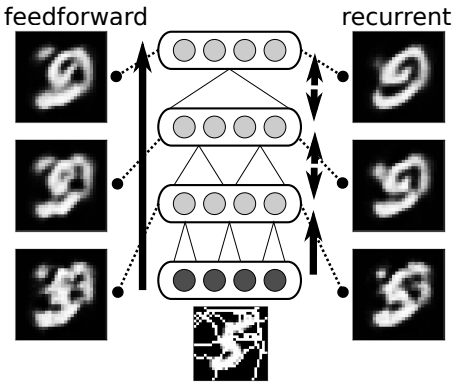
Figure 6.3 shows example decoded hidden states for the remaining two data sets, *shapes+clutter* and *MNIST+clutter*. For the former, in the RRF-DBM recurrent processing again made it possible to better retrieve the shape. Unlike in the case of *multi-shapes*, the plain DBM now however completely failed to recover the shape among the clutter. We found that with the clutter background, the input failed to activate the hidden layers sufficiently. With even first layer hidden units in the plain DBM having global receptive fields, the clutter images were too much of a mismatch to what the weights had learned in training. For the RRF-DBM on the other hand, the localised receptive fields meant that at least some hidden units in the first hidden layer could assume meaningful representations (where the shape was located in the image). Together with the tendency of higher layers to overcome undesired input, this was sufficient to realise the attentional selection. Although one should be cautious to generalise this finding to other contexts, it seems plausible that a restricted receptive field architecture makes the DBM inherently



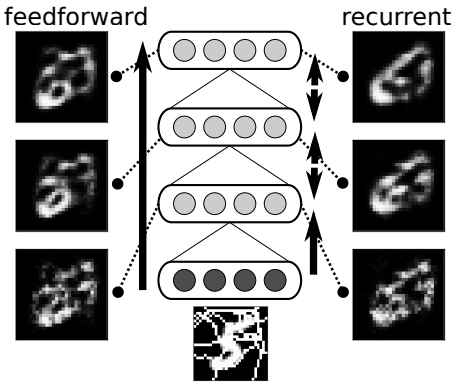
(a) RRF-DBM on *shapes+clutter*.



(b) Plain DBM on *shapes+clutter*.



(c) RRF-DBM on *MNIST+clutter*.



(d) Plain DBM on *MNIST+clutter*.

Figure 6.3: Further examples for the other two data sets. *(a)*: RRF-DBM on a *shapes+clutter* image. Again recurrent processing improved performance over the initial FF-sweep (although note that in this example, the category of the shape was actually misrepresented as triangle rather than square). *(b)*: the plain DBM now completely failed to recover the object, even with recurrent processing. Apparently the global receptive fields made it difficult to cope with the noise. *(c+d)*: for *MNIST+clutter*, neither RRF-DBM nor plain DBM performed that well, with or without recurrent processing. There was a tendency to 'hallucinate' image content as much as there was one to suppress the clutter, which could lead to the wrong internal interpretation being assumed.

more robust against noise unseen in training (also reported for deep belief nets by Tang & Eliasmith, 2010).

Finally, for *MNIST+clutter* (also Figure 6.3), neither RRF-DBM nor plain DBM performed that well, and recurrent processing gave little improvement over the FF-sweep. Examining the reasons for this, as well as suggesting possible mechanisms to address them, will be the subject of Section 6.4.

6.3.2 Quantitative analysis

To quantify how the models recovered individual objects in the images, classification and reconstruction errors were computed from the top layer's decoded states as explained in Section 6.2.2. In particular, for the *multi-shapes* images, which contained several candidate shapes, the 'attended object' was taken to be whichever one was reconstructed best.

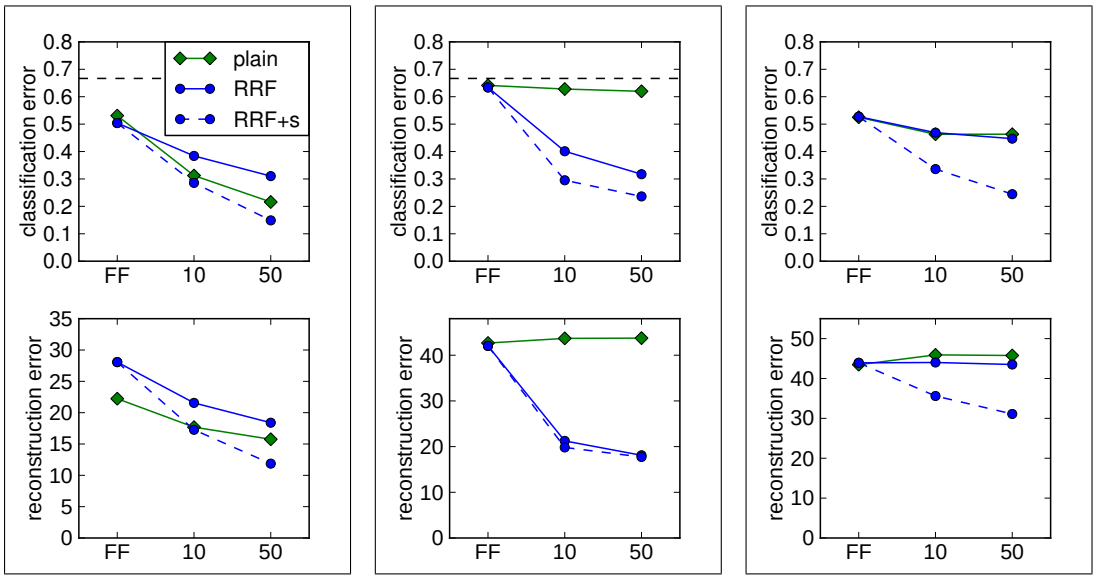
Results are displayed in Figure 6.4, overall confirming our analysis of the examples in the last section. For the *multi-shapes* set, the errors were rather high after the FF sweep, but dropped profoundly after subsequent recurrent processing cycles (e.g. classification error dropped from about 50% to about 20% for the plain DBM). This was true for both plain and RRF-DBM, the latter performing somewhat worse. For the noisy *shapes+clutter* set, performance was even worse after the FF sweep, with classification near chance. For the RRF-DBM, recurrent processing again helped greatly. Conversely, the plain DBM essentially failed to recover the shape among the clutter.

The results confirm that performance was mediocre for *MNIST+clutter* (albeit better than chance), with recurrent processing doing little to improve it. This will be discussed in the next section.

6.4 Experiments: top-down suppression on sparse representations

Recurrent processing did not help attentional perception for *MNIST+clutter*. In addressing the underlying problem we can further clarify the issue of attentional processing in the architecture.

Essentially, we treated additional image content as noise, and attentional selection was an emergent property of inference in a model that had learnt to represent individual objects only. The recurrent interactions in effect enabled the higher layers to override



(a) *multi-shapes*. (b) *shapes+clutter*. (c) *MNIST+clutter*.

Figure 6.4: Classification and reconstruction errors from top layer states for the three test sets. In each figure, scores are plotted for the plain DBM, RRF-DBM, and RRF-DBM with attentional suppression (will be explained in Section 6.4), taken after the FF-sweep and after 10 or 50 subsequent recurrent cycles. Dashed lines denote chance classification error (0.9 for MNIST, not shown). *(a)*: for the *multi-shapes* set, recurrent processing improved performance markedly. *(b):* for *shapes+clutter*, the restricted receptive fields of the RRF-DBM were moreover necessary to retrieve the shape. *(c)*: for *MNIST+clutter*, performance was better than chance but not that good overall, with little improvement due to recurrent processing. *With suppression*: for all data sets, suppression during recurrent processing enhanced performance further. It was essential for improving on the FF-sweep in the case of *MNIST+clutter*.

image content according to what they preferred to represent. In the case of the simple toy objects in the shapes data sets, this led to a suppression of other image content. On the other hand, the MNIST digits are much more variable in appearance than the simple toy shapes. When presented with, for example, a digit 5 among clutter, the model would ideally override the image representation in lower layers as to suppress the clutter. However, another way of reconciling the higher layers' expectation with the image could be to 'hallucinate' additional clutter to make the 5 into a 9 (cf. Figure 6.3c). Suppression or imagination of image content are equally possible. Similarly, the energy landscape learned from MNIST is likely to be more complex than that learned from the shapes, and thus supports more spurious states that might be more compatible with the cluttered images than the actual training data.

This thus brings up the conceptual issue of how the unattended content in an image should be treated in the probabilistic model formulation of the DBM, ideally in some principled fashion. We will address this question in the discussion (Section 6.5). For the time being, we take a more heuristic approach that views the DBM as a neural network, asking what neural mechanisms would make attentional processing more effective in the DBM. The resulting insights will be helpful for approaching the underlying conceptual problems as well.

6.4.1 The role of sparsity

A key issue is then that in the approach so far, recurrent processing in the model realises both a notion of Bayesian top-down predictions, where higher layers *change* the representations in lower layers according to what they expect, and attentional processing, where they *select* among various represented content. We need to tease apart these aspects so that attentional top-down selection can specifically increase the signal-to-noise ratio (signal being what is being attended to) without necessarily changing the content of the signal qualitatively, suppressing represented information related to the noise without 'hallucinating' additional content. To match the neurological effects of biological attention, this suppression of information should moreover correspond to actual suppression of activity of those neurons that do no represent the attended image content (and/or an enhancement of the activity of those neurons that do).

Two aspects of a standard DBM need to be considered to that end: first, in a completely distributed representation, where the image is encoded in the whole state vector $\mathbf{x}$, it is not clear how $\mathbf{x}$ is to be modified to achieve a suppression of image information

localised to a certain part of image space. In the RRF-DBM, this is solved by virtue of using the localised receptive fields, ensuring that units in lower layers only encode local information. The second issue is that switching an individual unit off (or on) does not necessarily correspond to suppressing information: for example, the unit could have inhibitory weights to the visible units representing the image. Indeed, we observed that for RRF-DBMs initialised with zero mean weights and biases, the learned representations were such that units tended to turn *off* when one of the shapes/digits was in their receptive fields.

Overcoming the second issue is why we initialised the unit biases to negative values at the beginning of training (Section 6.2.2). This lead to a breaking of symmetry between units being on and off, and particularly to units being only sparsely activated throughout training. Intuitively, they thus learned representations where they only turn on if something 'out of the ordinary' happens. In that sense, a unit conveys much more information by being on than by being off, and suppression of a unit can indeed be seen as effecting a suppression of represented information. With negatively initialised biases, units were indeed found to only turn on when some object (part) was in their receptive fields.

In this light, we note that the simple binary images we used in training might be somewhat less simplistic than they appear at first glance, as these images could themselves be regarded as proxies for a sparse representation of sensory input at an early stage of processing (having a high ratio of pixels being off to on).[7] We explored this further in our model extensions to be discussed in the end.

6.4.2 A suppressive mechanism for attentional selection

With such sparse representations established, we employed a simple heuristic suppressive mechanism to enhance the attentional processing, allowing higher layers to exert more influence on lower ones wherever their input to the lower neurons is suppressive, i.e. less than zero.

For a hidden unit i in intermediate hidden layer k, $k \in \{1,2\}$, we define the top-down input $T_i^{(k)}$ and the bottom-up input $B_i^{(k)}$ as

$$T_i^{(k)} := \sum_m w_{im}^{(k)} x_m^{(k+1)} + t_i^{(k)}, \tag{6.1}$$

$$B_i^{(k)} := \sum_l w_{li}^{(k-1)} x_l^{(k-1)} + b_i^{(k)}. \tag{6.2}$$

[7] This was also commented on in the chapter on hallucinations, Section 4.3.2

Here, $b_i^{(k)}$ and $t_i^{(k)}$ are the two bias variables for that unit, which originate from training the DBM layer-wise (in the process of which layer k is part of two RBMs). Normally these would be merged into a single parameter, but we do not do so here as they can be seen to contribute separately to bottom-up and top-down input. The probability for the unit to switch on is thus

$$P(x_i^{(k)} = 1|\mathbf{x}^{(k-1)}, \mathbf{x}^{(k+1)}) = \frac{1}{1+\exp(-B_i^{(k)} - T_i^{(k)})}. \tag{6.3}$$

The additional suppressive mechanism was then implemented simply by multiplying top-down input to a unit by a factor $\zeta^{(k)}$ ($\geq 1.$) whenever it was less than zero. This factor was set by hand for each layer. Writing the modified top-down input as $\tilde{T}_i^{(k)}$, the new activation rule for a hidden unit was thus now

$$P(x_i^{(k)} = 1|\mathbf{x}^{(k-1)}, \mathbf{x}^{(k+1)}) = \frac{1}{1+\exp(-B_i^{(k)} - \tilde{T}_i^{(k)})}, \tag{6.4}$$

with

$$\tilde{T}_i^{(k)} = \begin{cases} \zeta^{(k)} T_i^{(k)} & \text{if } T_i^{(k)} < 0 \\ T_i^{(k)} & \text{otherwise.} \end{cases} \tag{6.5}$$

Thus, this mechanism allowed a layer to suppress image content represented in a lower layer if it did not match its own representation, and to do so in a precise manner using top-down predictions. Again we emphasise that this merely a heuristic mechanism used to examine how how top-down predictions could in principle be utilised for attentional selection, and what the implications are for distributed representations.

6.4.3 Simulation results

The RRF-DBM experiments were repeated with the suppressive mechanism active in the intermediate hidden layers (also displayed in Figure 6.4). The performance increased in all cases. Particularly, for *MNIST+clutter*, recurrent processing with suppression now improved the scores markedly over the initial FF sweep.

In all cases, we tried different values for the suppression factor $\zeta^{(k)}$ for both intermediate hidden layers independently (values ranged from 1 to 6). The results reported are for the best values.[8] Notably, different settings worked well for the different data sets,

[8]For each test data set, the $\zeta^{(k)}$ values for the two intermediate hidden layers were chosen so as to minimise classification error after 50 recurrent cycles. A grid search with a step size of 1.0 was used, with $1.0 \leq \zeta^{(k)} \leq 6.0$, where $\zeta^{(k)} = 1.0$ corresponds to no suppression. The best values were: *multi-shapes*, $\zeta^{(1)} = 1.0$, $\zeta^{(2)} = 2.0$; *shapes+clutter*, $\zeta^{(1)} = 3.0$, $\zeta^{(2)} = 1.0$; *MNIST+clutter*, $\zeta^{(1)} = 6.0$, $\zeta^{(2)} = 1.0$.

presumably relating to how the receptive field sizes (which differed in the two hidden layers) interacted with the image content. For *multi-shapes*, enhancing suppression was beneficial in either of the hidden layers. For *shapes+clutter* and *MNIST+clutter* on the other hand, suppression was only effective in the first hidden layer. The reasons for this might differ in the two cases.

For *MNIST+clutter*, the digits filled relatively large regions of the images. Hence, suppression of small regions of clutter was more beneficial, and this required suppression of units with smaller receptive fields, i.e. units in the first hidden layer. For *shapes+clutter* on the other hand, the shape to be attended was relatively small, but localising it in the first place could actually involve some form of search: by inspection we observed that for some images, it took the internal model some time to converge onto the right location (note that in contrast for *multi-shapes*, initial coarse localisation is not difficult as all image content consists of potential candidates for attention). Thus, suppressing higher layer units and thus large regions of the image space early on in the process could be detrimental. More detailed analysis would be necessary to confirm our hypothetical explanations.

6.4.4 Spatial vs. object-based attention

In what sense is the attentional selection we modelled 'object-based'? The recurrent processing in the DBM and with it our additional suppressive mechanism operate in the hidden representations in a flexible fashion, and are not confined to spatially contiguous regions nor to single feature dimensions (which would correspond to forms of spatial and feature-based attention, respectively). Rather, they act on the basis of whatever forms stable representations in higher layers. They thus can be regarded as object-based as long as the predictions from the higher layers are object-based in some sense. In our model, this was arguably the case due to the training procedure used. It should be noted that biologically, the distinction between spatial and object-based attention is not always clear, especially if objects are spatially separated and the 'spotlight' of spatial attention is allowed to change shape to fit the outlines of an object (Scholl, 2001; see e.g. the model of Fazl et al., 2009).

With the topographic sparse representations in the RRF-DBM in place, we also briefly explored whether we could model spatial attention by applying a suppressive spotlight directly in the hidden layers of the model. To this end, we used a roughly Gaussian shaped spotlight acting as additional negative inputs to the hidden units,

suppressing neuronal representations away from a focus of interest. We tested whether this spatial spotlight could be employed to control which shape was being attended in the *multi-shapes* images, by applying it in the hidden layers such that it would be topographically centred on a randomly chosen shape in each image (in the brain, the source of the spatial attention signal is likely to be found in parietal or prefrontal cortex, e.g. Serences & Yantis, 2006).

We found that this method did indeed work. For example, classification error w.r.t. the selected shape after 50 cycles was about 18%, which is comparable to the scores reported in Figure 6.4a (plain DBM 22%, RRF-DBM + suppression 15%) for when the model could select any of the shapes freely.

6.5 Discussion

Bringing together generative models and attentional processing could be fruitful for both neuroscience and machine learning. From a biological point of view, generative models might not only offer insights into how the brain could learn about sensory input in an unsupervised fashion, and how top-down processing could be understood as higher stages providing expectations to lower ones, but possibly also into how these rich top-down predictions could be used to realise selective attentional processing.

For machine learning, attention suggests taking a perspective that considers probabilistic models and neural networks in the context of rich sensory contexts. Attentional processing would relate both to approximate inference algorithms that deal with the complexity of the involved tasks by focusing only on specific aspects at a time, and to the means by which sensory input is represented such that the relevant information becomes inherently accessible for control of behaviour. For Deep Learning approaches, this implies that the learned hidden representations would have to somehow relate to individual objects in complex visual scenes, such that for example information about segmentation is made available.

Although many open questions remain, our work might at least provide some ideas about how these aspects (approximate inference and behaviourally accessible representations) could come together: in a deep architecture where higher layers are by construction concerned with only one object at a time (or possibly, where attended object and the rest of the scene are separated somehow), not only should inference become more tractable, but the representations there should be inherently behaviourally useful in as far as that object is concerned. For example, in our work this includes localising

the object in the image by projecting the high-level representations back through the hierarchy. Similarly, the object-specific representations might be more easily interfaced with other modules and functions of a larger system, concerned e.g. with planning and action. As Duncan et al. (1997) writes in the biological context, "directing attention to a selected object makes its different features available together for control of behaviour and verbal report."

Notably, two recent machine learning papers used Boltzmann machines for attentional processing, but in a different sense (Larochelle & Hinton, 2010; Bazzani et al., 2011): both employed a 'fovea' of limited extent relative to the image, and were concerned with where it should be directed iteratively over time, and how information should be integrated across several such 'saccades'. This thus differs from our topic, which is internal attentional processing for a fixed sensory input, reflecting the distinction between covert attention in our case and overt attention in theirs. In future work, it would be very interesting to see how these different approaches could be combined.

With regards to biological approaches, there are several further computational models of hierarchical generative inference in the cortex which exhibit some attentional or attention-like effects (Murray & Kreutz-Delgado, 2007; George & Hawkins, 2009; Dura-Bernal et al., 2011). The focus in the cited studies was not attention however, and the demonstrated attentional effects were restricted to selected qualitative demonstrations. We will discuss these models in the discussion chapter in the context of approaches that are related to our model in general terms beyond individual perceptual phenomena (specifically in Section 7.1.5).

Lastly, a further, non-hierarchical approach related to ours is that presented by Rao (1999). A Kalman filter based generative model is used to learn about and recognise objects. Similarly to the predictive coding model of Rao & Ballard (1999) (also to be discussed in the last chapter), the internal model makes top-down predictions and computes residual errors from them. As in our model, training occurs on individual objects, and thus the predictions of the model can be used to implement attentional processing to segment objects from clutter. To this end, the predictions are made robust so that large deviations can effectively be ignored as outliers. Unlike our approach, Rao also uses grey-scale images of full objects. However, in all attentional simulations, his model was trained only on a handful of images, and attentional processing corresponded to exact recall of one of the latter. Moreover, aside from an additional noise process and the outlier aspect, the generation of images is a purely linear combination of 'causes', i.e. learned basis vectors. Indeed, with 5 training images, 5 basis vectors were

used, meaning that the internal model essentially just memorises (linearly recombined) training images.

Thus, while Rao's model is quite similar in spirit to ours, it is missing not only the hierarchical aspect, but is also likely in its current form not applicable outside of this specific demonstration of principle, both w.r.t. computer vision and biological questions. Nevertheless, its mechanism of making predictions robust and ignoring non-attended content as outliers is interesting. Moreover, computing residual errors allows for obtaining an explicit segmentation mask, which can then be used for example to switch attention to other image regions by inverting this mask. We describe some related mechanisms in our framework in the context of preliminary work below.

There are many more issues surrounding our exploratory work here. For instance, an analysis of the dynamics of neuronal activity in the model and how they compare with experimental findings would be beneficial to substantiating our work. In the remainder of this chapter, we address the following further issues: how hierarchical representations in the DBM could be made more powerful and what this would imply for attentional processing; how object-specific representations could be learned in the first place, and the role of motion processing; and, how a deep architecture could be designed that could deal with non-attended image content effectively and in a more principled fashion.

6.5.1 Invariant representations and attentional processing

One potential reason for why attentional processing or binding might be necessary is that precise location information about objects might be "lost" (Tsotsos et al., 2008) somehow in the ventral stream. In part, this might happen due to representations in higher cortical areas such as IT being spatially invariant, as has been assumed by models (e.g. Chikkerur et al., 2010). Location information can be reassociated with these representations by coupling them by some means with other areas where spatial information is more explicit, such as early visual areas or parietal cortex. This would allow for a localisation of the object (Tsotsos et al., 2008), and could itself play a role in realising the attentional binding across distributed feature maps (Treisman, 1996).

In the standard DBM, there is no mechanism that would encourage spatially invariant representations. In fact, in our model there is in a sense no loss of location information at all, as our decoding procedure shows that each pattern of activity in the topmost hidden layer corresponds to a detailed instantiation of the object in the image. One natural way to extend our model to explore attentional processing for more

invariant representations would be to use the convolutional model of Lee et al. (2009), which combines aspects of Boltzmann machines and feedforward convolutional neural networks (LeCun & Bengio, 1995). The latter are closely related to the HMAX model of Riesenhuber & Poggio (1999), a standard biological model of invariant representations in the cortex. We have made some initial attempts with this model, and continuing them could be a fruitful avenue for future work (also to be discussed in the final chapter, Section 7.2).

At the same time, it should be noted that the nature of invariant representation in cortical areas such as IT is a matter of debate. As discussed by Desimone & Duncan (1995) in the context of their biased competition account of attention, neurons in IT do have receptive fields that extend across large regions of the visual field (in the sense that they fire if a preferred stimulus falls within). At the same time, these regions can be inhomogeneous, and differ from neuron to neuron. Thus, it might be possible in principle to derive location information from the distributed representation of a population of neurons. Importantly however, while this information might be represented in principle, it is so only implicitly, and accessing it would require an actual decoder.

Thus, we would argue and clarify that hierarchical attentional processing might have an important function even if there is no loss of location information nor need for binding as such. For generative top-down pathways in general, the function would be to make implicit information explicit by transforming high-level representations into more interpretable low-level ones. For attentional processing, using similar pathways, the function would be to moreover make explicit which parts of the low-level representations relate to the currently attended object and which do not–which would be useful even if all of this were still implicit in the high-level representations. Indeed, even our decoding procedure (Section 3.5) relies on top-down processing to access what is represented in higher layers, because it uses the weights of (a copy of) the model to transform hidden representations back into images. In a sense, it employs the model of cortical processing, the DBM, as its own decoder. The brain cannot use such a separate decoder, and thus might rely on its own generative top-down connections to make this information more accessible.

6.5.2 Learning of object-specific representations from motion

We obtained a model that had learned representations specific to individual objects simply by training it on images containing only single objects. But how could the brain

form such representation initially when visual input is complex and cluttered most of the time?

While our developed brain can perceive objects among cluttered backgrounds, additional information might be necessary for it to initially learn to segregate objects from the rest of the scene, and there is evidence that motion could be crucial here. Ostrovsky et al. (2009) studied young Indian patients who had been blind from birth but then had their eyesight restored. Over the course of several months, they partially learned to see. At an early stage of recovery, the patients had great trouble to segment and recognise static objects (e.g. drawings of geometric shapes similar to our data sets) if there was background clutter or if objects overlapped. However, their performance greatly improved if objects moved independently. At a later stage, patients could then also handle static images, and there was further evidence that they had learned to recognise especially those classes of every-day objects that would often be encountered in motion. The authors conclude that "these results suggest that motion information plays a fundamental role in organizing early visual experience [...]".

This behavioural evidence matches various findings about motion processing area MT in the cortex. In development, it matures early on together with primary sensory areas (Bourne & Rosa, 2006). MT plays a role in segmenting figures from background, providing feedback to early visual area, possibly suppressing background regions (Likova & Tyler, 2008). And, it receives strong, fast visual input through subcortical routes (e.g. through the superior colliculus) in parallel to the usual thalamic pathway (Lamme & Roelfsema, 2000; Hupé et al., 2001). Feedback from MT to e.g. V1 might modulate early visual cortex even as or before visual input arrives (Hupé et al., 2001), and it can via V1 even reach thalamic inputs in a way that could relate to gating, attention, and segmentation (Sillito et al., 2006, see the latter for a review). MT cells also respond to the depth of a stimulus in visual space, another cue that could serve to segregate objects in a visual scene (op. cit.). Hence, MT and motion processing might be key for learning about objects in the visual system, and computational models would need to reflect that.

Finally, it should also be noted that there are various subcortical pathways implicated in attentional processing (Shipp, 2004). For example, the pulvinar is part of the thalamus, but it is actually driven not by the retina but input from visual cortex. Thus, "as its total output returns to the cortex, the pulvinar offers a good route for indirect transcortical communication" (Shipp, 2004). The existence of such parallel pathways

suggests that modelling attentional processing, or cortical feedback in general, with a single unitary hierarchical architecture might not be sufficient.

6.5.3 Towards novel deep learning architectures for attention

How could an attentional model be designed that can deal with challenging data more effectively, in a more principled way? In this final section, we report on some of our attempts in that direction. While some significant effort went into them, they were aborted at some point, ultimately because no improvement in performance over the earlier models was achieved, and because the relationship of the new model to processing in the brain was not clear. We thus keep the discussion rather brief.

For motivating the approach, some inspiration can come from the machine learning model of Titsias & Williams (2006). Though the latter is not framed in terms of attention at all (nor does it use deep learning or neural networks), the underlying concepts are relevant to our perspective on attentional processing, and it furthermore fits with our argument above concerning the importance of motion for learning about objects. In their study, Titsias and Williams were concerned with video sequences containing several moving objects. The goal was to learn without further supervision the object appearances, segmentation masks, and transformations, using a generative probabilistic model. Because inference in the full model was intractable, they utilised an approximate algorithm that learned sequentially about the objects in the scene, *one object at at time*. To this end, the algorithm represented the current object and background in a 'robustified' manner so that other objects could effectively be ignored as outliers.

Thus, their model is reminiscent of attentional theories that involve processing only one object at a time. Moreover, it suggests an alternative approach to our model presented in this chapter. Rather than having a hierarchy that ultimately attempts to only represent the attended object, and to suppress the rest of the image, one could use a model where attended object and background are segregated, but the latter is still represented. To keep the model in line with ideas of attention, most processing resources would however be dedicated to the attentional foreground. Moreover, fore- and background here would not be determined solely by the image content itself, but rather be assigned dynamically depending on the focus of attention, even if the input is static. At any point in time, the 'background' could indeed contain many other objects that could soon become the focus of attention (Figure 6.5).

Another way to frame dealing with the background would be in terms of an outlier

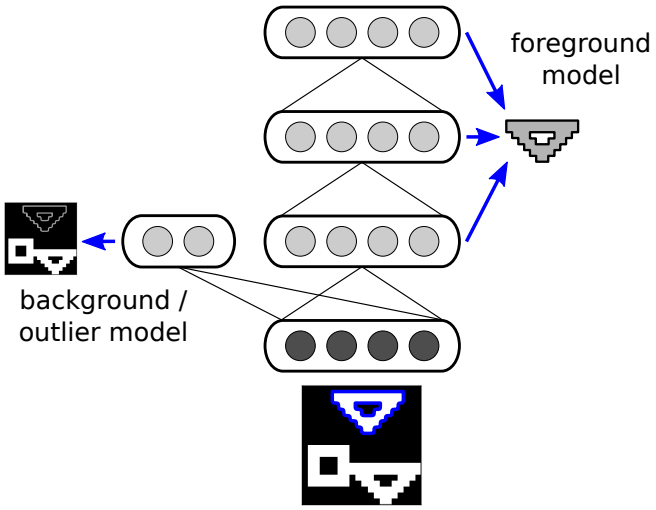


Figure 6.5: Example of how an alternative model architecture for attentional processing might look like. Most processing resources are dedicated to the object that is currently being attended, but the rest of the image is still being represented as well. Attention dynamically alters what constitutes attentional 'foreground' or 'background' in the static image. The segmentation might also be influenced by additional clues such as motion or depth.

model. Its representation only needs to be as good as necessary to allow the foreground model to ignore it. On the other hand, the less detailed background representation could also serve as context for the foreground inference, e.g. on a higher level correspond to the gist of the scene. Such an organisation would fit for example with the attention theory of Rensink (2000). In any case, it might also be possible to conceptualise the dynamic attentional separation of foreground and background as greedy approximate inference in a model of the whole visual scene, along the lines of Titsias & Williams (2006).

To translate these notions into a Deep Learning framework, we experimented with a version of restricted Boltzmann machine (RBM, which could be the first pair of layers of a DBM) where the hidden units and weights were separated into two sets, $\mathbf{W}', \mathbf{h}$' and $\mathbf{W}, \mathbf{h}$, corresponding to separate foreground and background models, respectively. An additional set of real-valued 'attentional' units $\mathbf{a}$ ($a_i \in [0,1]$) were meant to allow for gradual assignment or interpolation between fore- and background on a pixel by pixel basis. Inferring the states of theses units would also yield an explicit segmentation mask of attended foreground and non-attended background. The energy function was

$$E(\mathbf{v},\mathbf{h},\mathbf{a}) = -(\sum_{i,j} a_i v_i W'_{ij} h'_j + (1-a_i) v_i W_{ij} h_j) \tag{6.6}$$

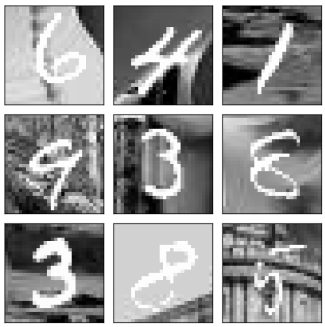


Figure 6.6: Example images from a variation of the MNIST data set containing digits on background images, after Larochelle et al. (2007), which we used for further attentional modelling.

(biases omitted). The idea was to give privileged resources to the foreground, by using more foreground hidden units, making only them part of a deep model, or similar. An alternative formulation removed the restriction of the fore- and background hidden units being strictly separated, but kept two sets of separate weights:

$$E(\mathbf{v},\mathbf{h},\mathbf{a}) = -(\sum_{i,j} v_i(a_i W^a_{ij} + W_{ij})h_j). \tag{6.7}$$

In that case, the attentional units could be seen as modulating the weights. With $\mathbf{W^a} := \mathbf{W'} - \mathbf{W}$ and enforcing that hidden units are only connected to the visible units via either $\mathbf{W'}$ or $\mathbf{W}$, Eq. 6.7 reduces to Eq. 6.6.

We trained and tested different models on the data sets presented earlier and on variations of MNIST as introduced by Larochelle et al. (2007), using MNIST digits on background consisting of images or random noise (Figure 6.6). As in that study, we treated the real-valued images as probabilities for the binary visible units. For simplification, we assumed that the attentional segmentation mask was given during training but absent during testing (which was motivated by the fact that additional motion information might provide a rough segmentation then).[9]

Unfortunately, for our data sets used earlier, performance was not satisfying and did in general not improve on what had been achieved with the original models as presented earlier in this chapter. For example, for *shapes+clutter*, no good background model could be learned. For the more interesting variations of MNIST, we found that treating the real-valued images as probabilities was problematic. The background model learned a white image and the foreground model digits without background. With an

[9]This required recreating the variations of MNIST from the components, as otherwise no segmentation masks would have been available for training.

input image given (but sampling the attention units), the attention units then actually interpolated these two images to roughly match the greyscale values of the background pixels, rather than properly assigning them to the background.

Because we wanted to keep binary visible units for now (also so that the same model could potentially be applied on the hidden layers as well, perhaps stacking these modules in a hierarchy), we instead tried using a model where the visible units did not represent the images themselves, but rather already a sparse representation thereof. To this end, we used a separate RBM with Gaussian visible units as a pre-processor, and then trained the attentional model on the hidden codes computed with the latter. This seemed to work in principle, but again, performance was not better than what we could achieve with standard versions of the DBM. Thus, we at some point stopped following this line of enquiry for the time being.

There are related avenues that could be explored, possibly considering other architectures than BMs as well. One idea would be to learn attentional processing by manipulating the cost function during training. Jaramillo & Pearlmutter (2007) trained an auto-encoder architecture, i.e. a multi-layer feedforward neural network with a bottleneck that learns to reconstruct its input, using an attentional signal that derived its meaning purely from a modulation of the cost function according to a spatial spotlight– essentially, the signal indicated what regions of an image input were 'important' to reconstruct, though the association of the signal with the spatial region had to be learned itself. After training, the authors found that the network could be directed to 'pay attention' to a spatial region using the attentional signal, encoding that region with higher fidelity, even though no explicit spatial spotlight was provided at that point.

It would be interesting to see if this approach could be used for more object-based attention, perhaps using an object specific segmentation mask during training (again with the motivation of ultimately using motion information, when available, to derive such a segmentation), and how it translates to a generative model such as the BM. The notion of a masked cost function is also implicit in our foreground/background model above, as given an attentional segmentation mask during training, the foreground model only needs to learn to generate the corresponding pixels as well. Then, when the mask is not given later, it needs to use what is has learned about the shape of objects to in particular decide which regions need to be attended. However, perhaps there are further ways of manipulating the cost function in a generative model, in the spirit of Jaramillo & Pearlmutter' approach. Lastly, it might also be fruitful to explore whether their approach could be applicable to neural networks with richer mechanisms than

auto-encoders, e.g. involving gating circuits (e.g. Hochreiter & Schmidhuber, 1997), and make it possible to learn effective forms of attentional processing there.

6.5.4 Conclusion

Our exploration showed that considering hierarchical generative models together with object-based attentional processing could be a promising avenue for future work. Many open questions remain, such as how more principled approaches could be formulated and how a closer connection to biology could be established. Of course, one cannot necessarily expect that processing in the brain maps naturally onto more principled machine learning methods, especially on the algorithmic or implementation level. Whether both biological realism and principled computations can be achieved within the same modelling framework remains to be seen.

Chapter 7

Discussion

In this thesis, we investigated the deep Boltzmann machine (DBM) as a model of generative processing and analysis by synthesis in the cortex. We modelled the emergence of visual hallucinations in Charles Bonnet syndrome, bistable perception, and a form of object-based attention. The DBM being both a probabilistic model and a neural network, we focused on aspects of approximate probabilistic inference and their interplay with neuronal mechanisms such as neuronal adaptation and homeostatic regulation of firing rate. Many issues surrounding our approach were already addressed throughout this thesis. In particular, we discussed our findings extensively in the respective results chapters, which included considering other models of the perceptual phenomena studied. We also provided a context for our work in terms of probabilistic approaches to cognition (Chapter 1), as well as elaborated on the technical background of the DBM and on questions relating to its biological interpretation, in Chapter 2 and Chapter 3, respectively.

In this last chapter, we focus on any remaining concerns. In Section 7.1, we review other computational models of generative processing in the cortex in general, in particular those that have not already been covered in the results chapters, and consider how they relate to our model. In Section 7.2, we suggest future lines of research in the DBM and related frameworks, also summarising what we have put forward to that end so far. Lastly, we conclude with some final remarks in Section 7.3.

7.1 Related models

Other computational models of the three perceptual phenomena were discussed earlier. Here, we give a brief overview over models of cortical representations and processing

that share some of the general computational principles that were the focus of our work, especially generative processing and hierarchical probabilistic inference. We contrast these models to the DBM, ask whether they could also serve to examine the same perceptual phenomena, and consider how the statements they make about cortical processing compare to those entailed by the DBM model. We restrict our discussion to computational models in the sense that they are mathematically formulated and involve simulation experiments. In our scheme for classifying probabilistic models presented in Section 1.2.3, these models are all internal, instrumental, and at an intermediate or low level of description.

7.1.1 Sparse coding and natural image statistics

A representative example of a class of generative approaches to cortical representation is the sparse coding model of Olshausen & Field (1996, 1997). There, natural images are encoded with an overcomplete set of basis functions that are linearly combined to reconstruct a given input image. The corresponding coefficients in that combination are taken to be the activities of neurons in primary visual cortex (V1). Olshausen & Field find that when they optimise this representation w.r.t. the reconstruction error while also encouraging sparse activations, the learned basis functions resemble Gabor filters (edge detectors) in a way thought to roughly capture receptive fields of neurons in V1. In technical terms, they define a probabilistic generative model with a sparse prior on the basis coefficients, and perform (local) MAP inference and learning of the parameters through an iterative procedure. The assumed goal of the cortex is to find an efficient code that exploits redundancy in the sensory input, and in doing so to discover structure and regularities in it (also Hyvärinen et al., 2009, Ch. 1). As the authors note, ultimately one would want to discover the "real causes of images (e.g. objects)", but their model is hampered to that end by the linearity assumption and lack of hierarchical structure.

The are several extensions to the sparse coding model as well as related approaches such as independent component analysis (ICA; see Hyvärinen et al., 2009, for review and details, on which we base this discussion). What they generally have in common is an approach to vision via models that capture the (low-level) statistics of natural images. These models are mostly evaluated in terms of the features they learn (Hyvärinen et al., 2009, p. 20), whether the underlying assumptions (such as independence) match natural image properties, and whether they achieve the stated goal of redundancy reduction.[1]

[1] Notably, Eichhorn et al. (2009) argue based on a quantitative analysis that orientation selectivity actually contributes only little to redundancy reduction.

They also consider activation effects that can be interpreted as competition between neurons, etc. They focus less on vision tasks like object recognition (at least not to begin with), perceptual phenomena, and the details of processing underlying cortical inference, and are mostly restricted to what would correspond to V1, lacking hierarchies, higher stages of processing, and effects of feedback (but see Hyvärinen et al., 2009, Ch.14, for multi-layer examples with feedback in the context of contour integration).

The model we employ, the DBM, learns about simplistic 'objects' and is hierarchical, even if it does not capture the rich computations of the cortical hierarchy. The sparse coding or ICA approaches are concerned with natural images (or image patches), and our model side-stepped many of the challenges they face by only modelling binary images. Still, for our purpose, idealised hierarchical representations of idealised visual objects had to suffice, as we were interested in aspects of hierarchical generative processing, object-based perception, etc. Thus, a shallow model of image patches would have been unsuitable. Looking towards the future, work on making Deep Learning approaches deal with proper images is ongoing, arguably with a larger focus on discovering higher-order structure in images, and sparsity plays a role (e.g. Lee et al., 2008; Nair & Hinton, 2009). Thus, we might see further connections with sparse coding models in the future. Our own intuition is that learning about objects might require stronger structural assumptions to be built into a model, e.g. related to object-based attention, and that it might be difficult to derive the powerful representations that are ultimately useful for complex behaviour purely from natural image statistics and efficient coding arguments.

7.1.2 Predictive coding

Another influential proposal is the predictive coding model of Rao & Ballard (1997, 1999). Their model consists of a hierarchy of neuronal layers where each layer predicts the activation patterns in the layer below (first layer being an image), and can be seen as a Kalman filter model. Predictive coding here is meant to specifically refer to a processing scheme were feedback signals carry predictions, and feedforward signals carry residual *errors* that remain when predictions are subtracted from the actual activations. The underlying assumption as stated by the authors is that sensory signals are generated "hierarchically via interacting physical causes (object attributes such as shape, texture and luminance)". The goal of the visual system then is to infer these causes, and on a longer time scales to learn the parameters of the generative model. Not surprisingly,

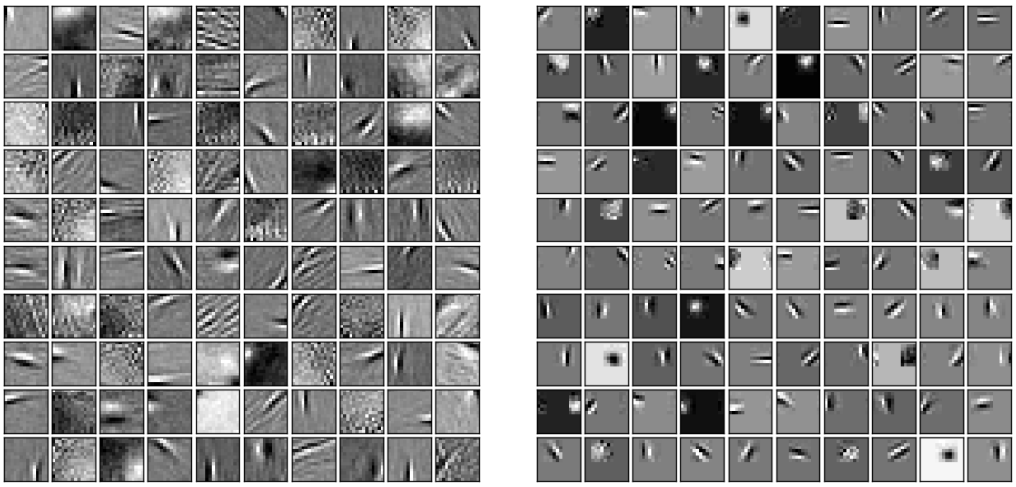
discovering such complex causes from natural images is difficult, and the model learns edge detectors and similar features. Similarly to the aforementioned sparse coding models, Rao & Ballard relate the model to cortex via the receptive fields it learns, as well as by explaining non-classical receptive field effects (such as 'end-stopping', to be discussed below) with the inference dynamics in the hierarchy, resulting from feedback. They also examine aspects of object recognition (Rao & Ballard, 1997).

As in our model, Rao & Ballard's model uses increasing receptive field sizes in the hierarchy. Whereas connections between layers are symmetric between layers in the DBM, clearly an idealisation of cortical connectivity, predictive coding assigns asymmetrical roles to feedforward and feedback connections, specifically in terms of propagation of error and prediction signals. Whether this scheme could be implemented in the cortex is being debated (e.g. Murray et al., 2004). To our knowledge, it has not been attempted to model perceptual phenomena, such as the ones we studied, with Rao & Ballard's hierarchical model, nor have its synthesis properties as a generative model (outside of inference) been examined. Rao & Ballard do address object-based attention, but only in a non-hierarchical version of the model (Rao, 1999; Rao & Ballard, 2004, Section 6.5). Also, Hohwy et al. (2008) give a predictive coding account of binocular rivalry (in the context of Friston's free-energy framework, see below), but they do not provide a computational model.

It would be interesting to explore perceptual phenomena with Rao & Ballard's model and compare results to our work. A model restricted to image patches would again be unsuitable. Rao & Ballard do test their model using greyscale images of whole individual objects (Rao & Ballard, 1997; also in the shallow version, Rao, 1999), though it should be noted that the object recognition tasks modelled effectively correspond to memorisation and recall of a small number of training images only.

7.1.3 Receptive fields and end-stopping

At this point we should address that we did not analyse receptive fields nor attempt to reproduce effects on neuronal responses such as end-stopping, both of which are commonly done for aforementioned approaches. The former we did not perform because our models were trained only on binary images, not natural image patches, thus they did not learn Gabor filters. We undertook preliminary experiments with natural images, Gaussian visible units, and a hidden layer subject to sparsity, and succeeded in learning such receptive fields in principle (Figure 7.1). An analysis of receptive fields in BMs in



(a) Unrestricted receptive fields. (b) Restricted receptive fields.

Figure 7.1: Receptive fields learned in the first hidden layers of two restricted Boltzmann machines, in some preliminary experiments with natural images as data (using Gaussian visible units, overcomplete hidden units, and sparsity). *(a)*: with weights between visible and hidden units unrestricted, some localised Gabor filters emerge, but some fields are global (which likely could be changed by further tuning the learning parameters). *(b)*: restricting receptive fields might encourage more Gabor filters.

biological terms has been undertaken however by others (Lee et al., 2008; Saxe et al., 2011). In particular, Saxe et al. performed an analysis of five different unsupervised learning algorithms all involving some notion of sparsity, including BMs and sparse coding, by comparing the receptive fields they learn to those measured experimentally in three different primary sensory cortices. All five algorithms were broadly consistent with experimental data, and results did not differ qualitatively between algorithms in any of the cases. The authors take this as evidence that learning across sensory cortices might be guided by the same (qualitative class of) algorithm. Their results might also show however that considering receptive field properties in low-level representations might only go so far in differentiating between different models, especially if all of them involve sparsity.

As for neuronal responses, end-stopping refers to a phenomenon where a neuron that responds maximally to a bar of a certain length in its receptive field has reduced response if the bar is made longer. Within the modelling approaches mentioned so far, there are two different explanations for this phenomenon. In predictive coding, the response of a low-level neuron corresponds to the error remaining after the higher layer

made its prediction about the image. If the latter has learned about long contours in natural images, then a longer bar will be better explained at the high-level, thus reducing the prediction error, i.e. activation, of the lower neuron, yielding end-stopping. In sparse coding on the other hand, end-stopping is a result of employing an overcomplete basis with a sparse prior on the activations, leading to units competing to explain the evidence (one unit 'explaining away' the evidence for others; Hyvärinen et al., 2009, Ch. 14). A long bar matches the preferences of several units, but sparsity encourages only few of them to be active at the same time, essentially corresponding to a mutual inhibition between them and thus producing end-stopping.

Could end-stopping be reproduced in our model? In the DBM, there is no representation of error signals as in predictive coding. If in the first hidden layer two units with small (classical) receptive fields respond to a long bar that a higher layer prefers to represent, then the feedback to the lower units would only increase their activation probability. We thus would have to rely on explaining away or competition effects like in sparse coding. In a DBM with only a single hidden layer (i.e. a RBM), due to the undirected connections, the hidden units are conditionally independent given the visible units, meaning that there is no interaction between them and thus no explaining away. Higher layers reintroduce dependencies between the first hidden layer units. However, again it is not clear to us at least intuitively why a higher layer should not support a long contour representation in the first layer rather than suppress it (unless long contours are actually rather rare in natural images and higher layers represent other kinds of large scale structures). Indeed, in the sparse coding model, end-stopping was observed in the shallow model. In the version with an additional layer (Hyvärinen et al., 2009, Ch. 14), feedback led to sharpening of a contour representation, i.e. suppression of distractors and some *enhancement* of responses along the contour. Presumably this countered the effect of end-stopping, though this is not discussed by the authors.

Thus, in both a hierarchical sparse coding model and the DBM, end-stopping might not be observed. This should be tested in the future. Possibly, additional mechanisms such as normalisation of population activity might be necessary to obtain end-stopping in a hierarchical model, outside a predictive coding scheme.

7.1.4 The free-energy principle

A different instantiation of a predictive coding model has also been put forward within the free-energy framework of Friston & Kiebel (2009); Friston (2009, 2010). The

scope of the latter is very broad, and the 'free-energy principle' is proposed to be not only a potential "unified brain theory" (Friston, 2010), but indeed to apply to "any biological system [...] from single-cell organisms to social networks" (Friston, 2009), and to subsume the Bayesian brain hypothesis (op. cit.). While we would be eager to scrutinise these claims further, this is beyond the scope of the discussion here. Rather, we do briefly comment on the concrete computational model that has been proposed for cortical inference as a special case within this larger framework.

This model of cortical processing is based on a generative model of data in terms of a hierarchical dynamic model in generalised coordinates of motion (Friston, 2008, 2009). For inference, in technical terms a variational mean-field approximation is used (explained towards the end of Section 2.2.2), assuming Gaussianity. Essentially, what this boils down to is to approximate the true posterior under the generative model with a unimodal distribution. Moreover, the resulting processing scheme also uses predictive coding with error signals being propagated feedforward and predictions feedback. Neuronally, this is suggested to map to the cortical hierarchy, and neuronal activity is thought to code for the sufficient statistics of the approximate posterior. This encoding scheme is thus an alternative to a population code of a full distribution or the sampling-based representation we employed throughout this work.

The hierarchical model is again suggested to be rather general in terms of what data it can analyse (Friston, 2008). However, the concrete biological examples focus on cases where there are few low-dimensional dynamical variables to be inferred, such as a toy example where parameters of a Lorenz attractor system are taken to be the control parameters for generating artificial 'birdsongs' (Friston & Kiebel, 2009). These parameters are then to be inferred by a 'synthetic bird'. To us at least, it is not obvious whether this approach would generalise to e.g. images, i.e. high-dimensional problems and/or cases where the structure of the true generative process of the data is not known to the system, or cannot be easily formulated even in principle (note that in the case of the 'synthetic bird', the true generative process as well as the perceptual inference are implemented by the same system[2]). Thus, it is not clear whether this computational model could be applied to the phenomena we studied in this work (again note the work by Hohwy et al., 2008 on binocular rivalry, which however lacks a computational model).

[2]It should be noted that bird brains do not actually have a cortex, possible commonalities notwithstanding (e.g. Kirsch et al., 2008). In so far as Frison's hierarchical model is meant as a concrete model of cortical processing, the usefulness of this example application might be limited.

7.1.5 Further hierarchical generative models of cortical vision

There are several further relatively recent computational models of hierarchical generative processing in the cortex that have or could potentially be employed to study some of the perceptual phenomena we examined. We can only briefly summarise these approaches here. Generally speaking, these models differ from ours in that they do not take a sampling-based perspective on inference, are often more complex, and do not cover several perceptual phenomena at the same time but focus more on individual ones, on aspects of object recognition and feature representations, and/or on biological realism.

Dean (2006) models learning of invariant features in visual cortex using a spatially extended hierarchical hidden Markov model. The underlying idea is to exploit that the causes of sensory input change more slowly than the input itself (as in slow feature analysis, Wiskott & Sejnowski, 2002), using 'inertial priors'. Inference and learning uses belief propagation and appears to be relatively complex. Dean focuses more on the principles of the algorithmic implementation, less on the neuronal interpretation, related perceptual phenomena, or performance. For comparison, the approach of Deco & Rolls (2004) is another example of a model focusing on learning invariances, and includes feedback processing for attention, but it is not formulated as a generative or probabilistic model.

Spratling (2011) presents a generative (non-probabilistic) model of predictive coding in the cortical hierarchy. The model is evaluated primarily in terms of whether it captures the response properties of cortical neurons, including effects of attention (Spratling, 2008), as well as in terms of the receptive fields it learns up to V2. The model is also shown to solve a simple artificial learning task known as the bars problem. While framed as generative model, its synthesis capabilities as such are not examined much. It also does not include stochasticity or adaptation. Thus, it is not clear whether Spratling's model could be used to model the phenomena we studied, but it would be interesting to explore, as its modelling of neuronal dynamics is more detailed, and the predictive coding scheme might lead to distinct effects.

Similarly to the aforementioned model of Dean, George & Hawkins (2009) propose a hierarchical probabilistic generative model of visual cortex consisting of a (tree structured) hierarchy of Markov chains (called Hierarchical Temporal Memory). Inference is performed using belief propagation. Though they do not discuss the biological plausibility of learning in this model, they go to great lengths to describe a putative mapping

of the mathematical entities of the theory to a complex neuronal circuit, which they argue to be situated in cortex according to a variety of known facts about cortical cell types and laminar organisation. The model is tested in object recognition tasks. Importantly, they also model recognition in clutter, including segmentation and localisation of an object via feedback. Although they do not frame it this way, this process appears to be quite similar to Tsotsos' Selective Tuning model of attention (Tsotsos, 2011b, Section 6.1.3), and thus similar to our own work. They also model subjective contour formation through feedback. As these perceptual aspects are not their primary focus, their discussion is somewhat brief.

Overall, it seems that George and Hawkins' approach could have quite some potential to model the perceptual phenomena we studied, in particular attention. Due to the complexity of their model, and due to the fact that they mostly apply it to custom data sets for which no performance comparisons exist, we find it somewhat difficult to judge the computational capabilities and theoretical implications of the model, and a detailed discussion of its proposed match to cortical circuits is not possible here. It might be noteworthy that in the most recent (commercial) instantiation of their algorithm as described in a white paper (Hawkins et al., 2011), they do not utilise feedback between cortical regions at all, for the time being. Thus, the generative or top-down aspect might not be essential to their approach, which would be a key distinction to models like the DBM where there cannot be learning without feedback.

Another example of a hierarchical generative model of biological vision is that of Murray & Kreutz-Delgado (2007). They start with a custom generative model that can be related to BMs but uses asymmetric weights and 'boltzmann-like' distributions. They then introduce an approximate variational mean-field model for inference and learning, and subsequently a dynamic neural network implementation of the latter. The model is trained on a custom image data set containing computer-generated objects, in a supervised fashion (providing object labels), using a version of backpropagation-through-time. The authors test the performance of the model on a number of visual inference tasks and analyse the evolution of neuronal activation patterns. Notably, these include segmentation from clutter and synthesis ('imagination'). In some cases, they even observe bistable behaviour of the network during synthesis. Thus, while they do not attempt to explicitly model hallucinations, attention, or bistable perception, there might be potential to do so in their framework. As they use supervision during training, their approach might not however fall into the class of unsupervised generative models

that could flexibly discover structure in data and thus possibly explain the versatility of the cortex, which motivated exploring the DBM as cortical model in this thesis.[3]

Finally, the recent work of Dura-Bernal et al. (2011) combines aspects of the well-known HMAX model of invariant feedforward object recognition (Riesenhuber & Poggio, 1999) with a Bayesian generative model using belief propagation for inference. In proofs of concept, they demonstrate illusory contour formation as well as attentional effects via feedback. This line of work seems promising and will hopefully be expanded in the future.

7.1.6 Conclusion on related work

There are thus several approaches modelling generative hierarchical processing in the cortex, to which our model can be compared. The DBM appears to be the only that combines the aspects of sampling-based probabilistic inference, neural processing, and unsupervised learning. Moreover, for our modelling applications, it was important that we could employ a strong form of internal synthesis (Eslami et al., 2012) of representations of structured sensory data, or 'objects' in particular, even if that meant using only simple binary images. This synthesis needed to occur in the absence of actual sensory input, and in a way that would not just correspond to recalling individual training images.[4]

Concrete differences in the algorithms aside, the reviewed approaches all differ somewhat from each other in their focus, the idealisations they make, and what aspects of cortical processing they neglect. For example, sparse coding and ICA approach vision from the perspective of natural image statistics, but focus less on visual tasks or the cortical implementation. Models such as the one of Dean (2006) on inertial priors, or the Hierarchical Temporal Memory of George & Hawkins (2009), explore general principles of cortical processing and to that end design novel, complex model architectures, but then also face the challenge of substantiating that these architectures can be related concretely to the cortical one (George & Hawkin's hypothetical but purely qualitative match of their model to cortical circuitry notwithstanding). Moreover, in so far as they make claims to capturing important principles of cortical processing

[3] Correspondingly, in the visual tasks they model that require a strong form of internal synthesis, namely imagination and segmentation from clutter, they need to provide object labels to the model, which is why they refer to the segmentation task as 'expectation-driven segmentation' (they do not mention the term attention at all).

[4] As noted above, the model of Murray & Kreutz-Delgado (2007) might have this capability, but it does not fall into the class of models trained without supervision.

and given that current machine learning and artificial intelligence still lacks behind in reproducing the capabilities of cortex, these models would ideally at least show potential for performing comparably well on e.g. computer vision tasks.

Similar considerations apply to our model. Due to being developed in the machine learning context, the DBM comes with a principled mathematical framework (that at least provides perspective when in practice many aspects of DBMs are based on approximate or even heuristic techniques), which connects it to various other approaches in machine learning and statistics. The interest in machine learning also means that models like the DBM are put to the test when it comes to solving actual challenging tasks. Though performance was not the focus of our study, such context can at least provide a sanity check that the assumed computational mechanisms for learning representations and inference are sensible. Still, Deep Learning has yet to fulfil its promise of delivering powerful solutions to problems such as those in vision when it comes to more than simplified problems. As further developments take place, we see further potential for biological modelling applications, as will be shown in the next section.

7.2 Future work

Several avenues of possible future work have been proposed throughout this thesis. We summarise them here and make additional suggestions for further biological modelling applications within the DBM and Deep Learning framework.

7.2.1 Rich deep architectures as cortical models

In machine learning, there are multiple ongoing developments that could enrich DBM-like models in ways that are biologically relevant, potentially making it possible to extend our modelling work so far or to apply the DBM to other biological phenomena. Often these developments are realised in the related deep belief nets or in other BM-based models, hence the theoretical challenges of translating them to the DBM would also need to be subject of future work. As a first example, Osindero & Hinton (2008) include lateral connections in a deep belief net, a feature of cortical circuits amiss in the DBM. The underlying motivation of their work is to develop hierarchical models where higher layers only need to give rough predictions, e.g. about where different parts of an object should be positioned approximately, with lower layers sorting out the details

via local interactions. Not having to learn detailed local correlations could also free up higher layers to learn more interesting structure in the data. Thus, exploring this further could offer means of making the hierarchical computations in DBMs richer. It could result in an architecture where higher representations are less concrete and thus possibly suitable to study the differences between mental imagery and hallucinations (to be addressed below). It might also allow modelling the dynamics of cortical computations in terms of feedforward, feedback, and lateral interactions more realistically, and suggest additional functionality for top-down processing e.g. in the context of attention (cf. Section 6.5.1).

Also promising along similar lines would be to model how invariant representations are formed in the hierarchy in a way that relates to biological models. Lee et al. (2009) introduced a convolutional BM-type model that includes aspects of convolutional neural networks (LeCun & Bengio, 1995), which in turn closely relate to the biological HMAX model of Riesenhuber & Poggio (1999). The latter is a standard model of how hierarchical invariant representations are formed in the cortex by utilising 'complex cells' that nonlinearly pool outputs from sets of 'simple cells' responding to similar image features. We indeed reimplemented Lee et al.'s architecture and performed initial experiments, but did not proceed very far due to initial technical challenges (not discussed here for brevity). Similarly, the BM model of Ranzato et al. (2010) attempts to better capture covariance structure in natural images, and due to the way the hidden units pool information, the authors compare them to complex cells. Ranzato et al.'s model is also an example of a *third-order* BM (Sejnowski, 1986), where a weight connects not two but three units. In particular, this can be interpreted as one unit modulating or gating the interaction between two other units. Such mechanisms could play a role in various biological contexts as well, such as gating of information for attentional processing or working memory. Neuronally, such gating could for example be implemented in dendrites, where propagation of distal synaptic activation to the cell soma can depend on further inputs being present along the dendrite (Jarsky et al., 2005). There is also evidence for gating between cortical layers controlled by inhibition (Tiesinga et al., 2008).

Another interesting issue is whether and how hidden layers in deep models could be separated into specialised modules that preferentially represent certain aspects of the data, to match the cortical organisation into specialised regions. This would also make it possible to test whether selective deprivation of a visual feature dimension leads to corresponding hallucinations (Section 4.4.1).

7.2.2 Modelling the synthesis of visual experience

Future work could explore further issues concerning how visual experience is accounted for in generative (not necessarily probabilistic) models of cortical processing. As we have argued in Chapter 4, complex hallucinations, and possibly dreams, can be seen as evidence that the brain is capable of a strong form of synthesis of internal representations. In particular (Section 4.4.4), the rich detailed content of hallucinatory imagery must be instantiated somehow in these representations. Similarly, in bistable perception caused by the Necker cube, there might be a lot of information in the percept that is inferred beyond what is determined by the actual image (such as implied depth ordering of the lines), information that is itself subject to the perceptual switch. Arguably, this might also speak to why it could be ultimately more elucidating to model this phenomenon with sampling in a rich high-dimensional representation of images, rather than a low-dimensional abstract variable (as in Sundareswara & Schrater, 2008). In either case, there could be further interesting modelling questions concerning the relation between internally synthesised representations and evoked experience, constraining what kind of internal model the brain might implement. Key here is that addressing the deeper philosophical issues surrounding the nature of conscious experience can be avoided. The fact that someone experiences (or, at least, reports) complex hallucinations necessarily implies that the entailed information must *somehow* be represented and addressable, in whatever form. This is a necessary condition, even if it is not clear what form of representation is *sufficient* to be accompanied by conscious experience. Thus, this information content should be reflected by a computational model.

An example question that could be addressed is why the experience associated with mental imagery appears to be much less rich than that of complex visual hallucinations or actual seeing. In a hierarchical generative model in the cortex, it seems plausible that this richness relates to extent of which internal representations are synthesised, in particular if representations in higher areas are more abstract according to some measure, and those in lower ones more concrete and detailed. The key aspect of a computational model would thus be that detailed information is lost by some means, or at least becomes inaccessible, in the feedforward processing from lower to higher areas, and that when generating from the model, this information would be gradually restored when going in the reverse direction. Hence, recent machine learning developments that have been discussed earlier in the context of enriching deep architectures could be highly relevant here. On the feedforward side, biologically plausible mechanisms such

as in-built translation invariance (complex cell pooling) could lead to some information not being represented in higher areas. On the generative side, this information loss could be inverted with mechanisms such as lateral interactions, so that higher layers only *seed* images in an approximate fashion, and lower arrange the details, e.g. by aligning edges (Osindero & Hinton, 2008; Hinton, 2010a). Then, lower areas really would be needed to realise all information entailed in rich perception, thus possibly explaining the perceptual difference between mental imagery, mostly constrained to higher areas, and system-wide vivid visual hallucinations.[5]

7.2.3 Rich perceptual, behavioural, and anatomical contexts

Future work could also be concerned with richer perceptual tasks and putting models into behaviourally interesting contexts that go beyond image recognition. For example, extending deep networks to videos and motion (e.g. Chen, 2010) might be key for learning about objects (Section 6.5.2) and invariances (e.g. Zou et al., 2011). Depth is another relevant sensory cue, which could be instrumental to modelling bistable perception for the Necker cube more realistically (see discussion and preliminary work in Section 5.5.5). Thus, recent work (Memisevic & Conrad, 2011) on learning about depth in BMs using stereo images could be considered here.

Taking as motivation attentional vision via eye movements, two recent papers on BMs were concerned with how information can be integrated as a limited 'fovea' is moved across an image or video sequence, and how a gaze strategy can be learned (Larochelle & Hinton, 2010; Bazzani et al., 2011). It could be fruitful to examine how plausible these approaches are from a biological perspective, and to combine them with covert attentional processing as modelled in our work. In the approach of Bazzani et al. (2011), the BM is part of a larger system tasked with identifying and tracking moving objects in videos, which for example also involves reinforcement learning. For future work, one might similarly consider a model like the DBM, taken as model of cortical representations, as it interacts with different kind of systems modelling other parts of the brain, such as the basal ganglia involved in action selection and reinforcement learning (e.g. Frank & Badre, 2011). Also, structures like the superior colliculus or the part of the thalamus called pulvinar are reciprocally connected with the cortex and thought to be

[5]This is not to say that mental imagery cannot lead to activation of even early areas. To the contrary, there is much evidence that it does, though to what extent is a subject of debate, as is whether activation of lower areas is a functional aspect of mental imagery or just an epiphenomenon (Kosslyn & Thompson, 2003). The difference between hallucinations and imagery could be a quantitative rather than a qualitative one. A computational model could help to address questions such as these.

important for attentional processing (Shipp, 2004). Understanding cortical processing especially in the context of complex behavioural tasks might thus be infeasible without taking into account the other brain systems it closely interacts with. In turn, a single homogeneous model architecture might be insufficient to solve such tasks.

Another example of an important system interacting with, and being closely related to, the cortex is the hippocampus,[6] and arguments have been made in connectionist terms why two separate systems with different types of knowledge representation might be necessary for learning (McClelland et al., 1995). Perhaps such reasoning also translates to the DBM. Of possible interest is the recent work of Salakhutdinov et al. (2011b). To realise learning from few examples, they present a model consisting of a DBM that learns a rich distributed feature space of images, and a nonparametric tree-structured model that sits on top of the DBM and learns a category hierarchy on that feature space. It is not clear whether their approach relates to the hippocampal-cortical system, but their work demonstrates that hybrid architectures consisting of DBMs and other types of models are feasible in principle and can be very effective in practice.

7.2.4 Learning in DBMs and the cortex

Taking the learning algorithms used in BMs and DBMs as models of biological learning was not subject of our work, though we made various suggestions to that end in Section 3.4 (see there for details) and thus see potential for interesting future work. The key ideas would be to relate CD (contrastive divergence) like learning to prediction error driven learning, free sampling from the model in the negative phase of the weight updates (e.g. during PCD) to dreams, and greedy layer-wise learning (as is common in Deep Learning approaches) to hierarchical development of cortical areas (Bourne & Rosa, 2006).

Relevant here might also be recent work by Zhou et al. (2012), who train autoencoders in an online fashion to adapt to a continuous stream of data, the distribution of which can also change over time. New feature units are constantly being added and old ones merged to keep the total number bounded. From a biological view, such online adaptation might be preferable to learning with a fixed training data set as is usually done in Deep Learning.

Finally, we should also note again that in our work, we only used the layer-wise pretraining of the DBM and never subsequent training of the full model. Seeing whether

[6] Technically, the hippocampus is itself 'cortex'. With the latter term we refer to specifically the cerebral neocortex.

the latter bears any effect on our results as presented in this thesis would be a sensible line of future enquiry.

7.2.5 Remaining issues

For sake of completeness, we summarise briefly any remaining possibilities for future work that were raised earlier in the thesis but have not been addressed above.

To expand our work on bistable perception, it would be very helpful to perform a detailed analysis of the interactions of stochasticity and neuronal adaptation (e.g. Shpiro et al., 2009), and to reproduce a key psychophysics experiment of Kang & Blake (2010) that showed that both adaptation and noise are necessary to account for the dynamics of binocular rivalry. Reproducing other findings could be attempted as well (Section 5.5.6). In the case of object-based attention, we have already undertaken preliminary work towards designing model architectures that could deal with more challenging data in a possibly more principled way (Section 6.5.3), modelling attentional selection as a dynamic assignment of image regions to attended foreground and non-attended background. Related issues not mentioned so far that could be addressed would be how occlusion should be handled, and the phenomena of amodal completion, i.e. implicit completion of occluded figures, and of modal completion (e.g. Murray et al., 2004), where foreground shapes are completed by filling in 'illusory contours'. The latter would be natural to model with internal synthesis in the DBM. Also, further work could focus more on 'artificial electrophysiology', i.e. on examining neuronal response properties in the model to account for phenomena such as end-stopping, as well as on extending the work of others on analysing receptive field properties (both Section 7.1.3).

Finally, an important issue is how models formulated at different levels of abstraction could be bridged. For example, the model of bistable perception of Sundareswara & Schrater (2008) is high-level and conceptual in that the sampling process underlying bistablity is formulated in terms of a variable explicitly characterising the orientation of a cube in an ambiguous image (Section 5.5.1). In our own model, sampling occurs in the implicit, high-dimensional, distributed hidden representations, and a different sampling algorithm is used. Whether these different models make conflicting predictions or whether they can be seen as compatible descriptions at different levels of abstraction remains to be examined. Similarly, it will be essential in the future to ask how high-level algorithms used in cognitive models, for example particle filters defined over category

variables (Sanborn et al., 2010), can be related and translated to more low-level models of neuronal representation and processing.

7.3 Conclusion

Reviews such as the one by Mesulam (1998) really reveal the wide range of functions the cortex is involved in, not only in perception, but also in motor control, abstract thought, language, and planning and decision making. 'Function-first' approaches (Griffiths et al., 2010) to understanding cognition, depending on context also framed as rational, ideal observer, or computational level models in Marr's sense, start by asking, how could a given problem posed by the environment be solved computationally, perhaps under the constraints imposed by biology. Given how many different functions cortex serves, such approaches will likely offer a multitude of different perspectives, different angles from which to shed light on cortical processing. Two challenges then are, to bring together these different perspectives, and to bridge levels of analysis from high-level functional descriptions to models of neuronal implementations. We believe that both these challenges will have to be addressed together: by devising mid-level, versatile computational models and algorithms that at least offer the potential to be connected to a variety of high-level functions, while at the same time being informed by the neuronal mechanisms in the cortex.

At the same time, a notion that there as a single algorithm that would underlie all of cortical processing might be naive. After all, cortical hardware had time to adapt during evolution, and anatomical specialisation that can be observed in different cortical regions could reflect qualitative changes in computational mechanisms as adaptation to diverse functional roles occurred. Nevertheless, to us it still seems worthwhile to ask whether a computational basis could be identified that would offer insights into why cortical adaptation to different functions was and is so successful–whether it is adaptation through learning or through evolution. In a sense, this is asking what the computational mechanisms of cortex were at an early stage of evolution before diversification took place (Kaas, 2011).

As others have argued before, we believe that the notion of a generative model in the cortex (and related concepts like analysis by synthesis, predictive coding, etc.) could offer one possible starting point. The key aspect of the resulting approaches to modelling is to see feedback or top-down processing at the core of what the cortex is about, rather than something that can be neglected in a first-order approximation

of cortical processing, starting with a feedforward model and then adding feedback back in later. The reason why generative processing is promising is that, if the right generative mechanisms were to be found, many diverse aspects of, or explanations for, cortical functions might follow naturally: unsupervised learning of an internal model from sensory data; using high level, prior knowledge to inform lower level and bottom-up processing (e.g. object and shape representations to guide segmentation); predicting future or unseen sensory data from current knowledge and beliefs; and, using generative mechanisms for other top-down aspects of perception such as attention. Similarly, it seems impossible to even consider non-perceptual synthetic processes such as language, motor control, and planning, without at least some form of generative component.

Unfortunately, coming up with good generative models, biological or not, of even modestly complex sensory data (handwritten digits?) has proven difficult. In this thesis, we took the approach of considering current relevant developments in machine learning, and demonstrated how one specific model, the DBM, could be employed as an idealisation of hypothetical generative processing in the cortex. We used this model to put forward several concrete hypotheses and theoretical insights regarding the mechanisms behind perceptual phenomena, such as: that internal synthesis in a generative model could, together with homeostatic regulation of neuronal activity, underlie the emergence of complex visual hallucinations in Charles Bonnet syndrome; that bistable perception and neuronal adaptation could be understood as aspects of sampling-based probabilistic inference; and, that generative processing could be a component of realising object-based attention.

Overall, our goal was to lend further credence to the hypothesis of generative processing in the cortex. Whether we have achieved this goal is for the reader to decide. In any case, the approach we took, namely considering methods (which are, at best, biologically inspired) from fields like statistics or machine learning as serious biological models, or at least as starting points to that end, is arguably quite common, even if it is not always framed as such. Examples are, arguably: connectionist neural networks; Rao and Ballard's predictive coding model, which is based on Kalman filters (Rao & Ballard, 1997); Friston's model of cortical processing that originated in approaches in statistical physics and machine learning (Friston & Stephan, 2007); and, the various sampling-based accounts of neuronal processing and cognitive inference (employing particle filtering, Sanborn et al., 2010, MCMC, Hoyer & Hyvärinen, 2003; Gershman et al., 2009b, etc.). We see further potential to that end in particular for Deep Learning models in machine learning (Bengio, 2009), as well for other neural networks that we

will not discuss here for the sake of brevity (e.g. Hinton et al., 2011; Hochreiter & Schmidhuber, 1997).

It is in the light of such considerations that we conclude with some final comments on the currently ongoing debate regarding Bayesian models of cognition (Sections 1.2). The subject of optimality has been one of the main points of contention in the recent debate (Bowers & Davis, 2012a; Griffiths et al., 2012; Bowers & Davis, 2012b). Based on reasoning elaborated on earlier, we would be careful with using the term 'optimal' outside of a well-characterised ideal observer analysis. In particular, optimal appears to sometimes refer to Bayesian inference being optimal *given* the underlying assumptions encapsulated in the model. But there is a danger of post-hoc declaring *any* perceptual process as Bayesian (Bowers & Davis, 2012a), and thus 'Bayes-optimal', because there is likely *some* Bayesian model in which that process implements inference, especially once one admits approximations. If it is not just optimal solutions to environmental challenges, what then are Bayesian approaches 'all about' (Bowers & Davis, 2012b; Jones & Love, 2011b)?

In our view, Bayesian approaches involve: a set of interrelated questions concerning how brains deal with uncertainty; theoretical means of deriving, in well-constrained circumstances, ideal observers; a set of theoretical tools to characterise people's perceptual and inductive biases; and, computational approaches that can serve as idealisations of perceptual inference, learning, and decision making, starting points and points of comparison for more realistic models. Because Bayesian models are often rather crude idealisations of actual cognitive processes, they might sometimes be justified as ideal observers or optimal solutions even when neither the problem to be solved nor what is optimal (or what is being optimised) can be properly quantified. In our view, Bayesian models of that sort might not need this justification, because *most* models we have of cognitive processes are crude–especially those at the forefront of our attempt to understand how the brain solves the most challenging aspects of cognition, such as vision, reasoning, or induction. Connectionist networks and Bayesian models are all crude idealisations, and they are all *wrong*, just in different ways (i.e. they are Galilean idealisations, Section 3.2). Just like Rutherford's planetary model or Thomson's 'plum pudding model' of the atom (Lakhtakia, 1996), these models are only stepping stones to further our understanding of the brain, an understanding that is still severely lacking at this point (the difference being, unlike with the atom, we might never arrive at a 'complete theory of the brain', due to the brain's complexity).

Even those researchers who see full Bayesian models primarily in the role of de-

scribing rational observers and optimal solutions appear to agree that ultimately, it is the approximate models, and the models that bridge levels of description down to neuronal processes, that are required to understand cognition (Griffiths et al., 2010; Sanborn et al., 2010; Griffiths et al., 2012). As the focus shifts towards these more realistic models, the question whether the original starting points, the full Bayesian models, represented optimal solutions, or just idealised solutions like many other function-first approaches, might become moot.

Bibliography

Aakerlund, L., & Hemmingsen, R. (1998). Neural networks as models of psychopathology. *Biological Psychiatry 43*(7), 471–482. Cited on page 97.

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science 9*(1), 147–169. Cited on pages 20, 50.

Anderson, B. (2011). There is no such thing as attention. *Frontiers in Psychology 2*. PMID: 21977019 PMCID: 3178817. Cited on page 145.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review 98*(3), 409–429. Cited on pages 13, 14.

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning 50*(1), 5–43. Cited on page 25.

Bartels, A. (2009). Visual perception: Converging mechanisms of attention, binding, and segmentation? *Current Biology 19*(7), R300–R302. Cited on page 149.

Bazzani, L., Freitas, N., Larochelle, H., Murino, V., & Ting, J.-A. (2011). Learning attentional policies for tracking and recognition in video with deep networks. In L. Getoor & T. Scheffer (Eds.), *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, New York, NY, USA, pp. 937–944. ACM. Cited on pages 168, 190.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning 2*(1), 1–127. Cited on pages 3, 4, 18, 20, 39, 45, 46, 100, 194.

Bengio, Y., & Delalleau, O. (2009). Justifying and generalizing contrastive divergence. *Neural Computation 21*(6), 1601–1621. Cited on pages 37, 52.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., & Bengio, Y. (2010). Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*. Oral Presentation. Cited on page iv.

Berkes, P., Orbán, G., Lengyel, M., & Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science 331*(6013), 83 –87. Cited on page 14.

Bialek, W., & DeWeese, M. (1995). Random switching and optimal processing in the perception of ambiguous signals. *Physical Review Letters 74*(15), 3077–3080. Cited on pages 106, 134.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer. Cited on pages 24, 51.

Blake, R. (1989). A neural theory of binocular rivalry. *Psychological Review 96*(1), 145–167. PMID: 2648445. Cited on page 106.

Bourne, J. A., & Rosa, M. G. (2006). Hierarchical development of the primate visual cortex, as revealed by neurofilament immunoreactivity: Early maturation of the middle temporal area (MT). *Cerebral Cortex 16*(3), 405 –414. Cited on pages 54, 171, 191.

Bowers, J. (2010a). More on grandmother cells and the biological implausibility of PDP models of cognition: a reply to Plaut and McClelland (2010) and Quian Quiroga and Kreiman (2010). *Psychological review 117*(1), 300–308. Cited on page 51.

Bowers, J. (2010b). Postscript: Some final thoughts on grandmother cells, distributed representations, and PDP models of cognition. *Psychological Review 117*(1), 306–308. Cited on page 51.

Bowers, J., & Davis, C. (2012a). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin 138*(3), 389–414. Cited on pages 7, 12, 14, 195.

Bowers, J., & Davis, C. (2012b). Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychological Bulletin 138*(3), 423–426. Cited on pages 7, 10, 12, 195.

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychological review 116*(1), 220–251. PMID: 19159155. Cited on page 50.

Breuleux, O., Bengio, Y., & Vincent, P. (2011). Quickly generating representative samples from an RBM-Derived process. *Neural Computation*, 1–16. Cited on pages 6, 38, 106, 107, 111, 112, 113, 116.

Bruce, N. D. B., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision 9*(3). Cited on page 145.

Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011). Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLoS Comput Biol 7*(11), e1002211. Cited on page 48.

Burke, W. (2002). The neural basis of Charles Bonnet hallucinations: a hypothesis. *Journal of Neurology, Neurosurgery & Psychiatry 73*(5), 535–541. Cited on pages 63, 64, 67, 70, 95, 98.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing 37*(1), 54–115. Cited on pages 2, 105.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research 51*(13), 1484–1525. Cited on pages 145, 147.

Chakravartty, A. (2011). Scientific realism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2011 ed.). Cited on page 11.

Chalk, M. (2012). *The role of expectations and attention in visual processing and perception*. PhD thesis, University of Edinburgh, Edinburgh, UK. Cited on pages 150, 151.

Chater, N., Goodman, N., Griffiths, T. L., Kemp, C., Oaksford, M., & Tenenbaum, J. B. (2011). The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science. *Behavioral and Brain Sciences 34*(04), 194–196. Cited on page 9.

Chen, B. (2010). *Deep learning of invariant spatio-temporal features from video*. MSc thesis, University of British Columbia, Vancouver, British Columbia, Canada. Cited on page 190.

Chikkerur, S., Serre, T., Tan, C., & Poggio, T. (2010). What and where: A Bayesian inference theory of attention. *Vision Research*. PMID: 20493206. Cited on pages 151, 152, 153, 169.

Collerton, D., Perry, E., & McKeith, I. (2005). Why people see things that are not there: A novel perception and attention deficit model for recurrent complex visual hallucinations. *Behavioral and Brain Sciences 28*(06), 737–757. Cited on pages 61, 63, 64, 65, 66, 90, 96, 148.

Corlett, P., Frith, C., & Fletcher, P. (2009). From drugs to deprivation: a Bayesian framework for understanding models of psychosis. *Psychopharmacology 206*(4), 515–530. Cited on pages 65, 97.

Courville, A., Bergstra, J., & Bengio, Y. (2011). Unsupervised models of images by spike-and-slab RBMs. In L. Getoor & T. Scheffer (Eds.), *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, New York, NY, USA, pp. 1145–1152. ACM. Cited on page 100.

Crick, F., & Mitchison, G. (1983). The function of dream sleep. *Nature 304*(5922), 111–114. PMID: 6866101. Cited on page 53.

Daw, N., & Courville, A. (2008). The pigeon as particle filter. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems 20*, pp. 369–376. Cambridge, MA: MIT Press. Cited on pages 16, 109, 136, 137.

Dayan, P. (1998). A hierarchical model of binocular rivalry. *Neural Computation 10*(5), 1119–1135. Cited on pages 106, 109, 122, 130, 134, 136, 139, 142.

Dayan, P., & Daw, N. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience 8*(4), 429–453. Cited on page 3.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation 7*(5), 889–904. Cited on pages 53, 100, 105, 156.

Dayan, P., & Solomon, J. A. (2010). Selective Bayes: attentional load and crowding. *Vision Research 50*(22), 2248–2260. Cited on page 150.

Dayan, P., & Zemel, R. (1999). Statistical models and sensory attention. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, Volume 2, pp. 1017 –1022 vol.2. Cited on page 150.

Dean, T. (2006). Learning invariant features using inertial priors. *Annals of Mathematics and Artificial Intelligence 47*(3), 223–250. Cited on pages 17, 184, 186.

Deco, G., & Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research 44*(6), 621–642. Cited on page 184.

Desai, N. S. (2003). Homeostatic plasticity in the CNS: synaptic and intrinsic forms. *Journal of Physiology-Paris 97*(4–6), 391–402. Cited on pages 67, 69.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience 18*(1), 193–222. Cited on pages 146, 148, 170.

Desjardins, G., Courville, A., Bengio, Y., Vincent, P., & Delalleau, O. (2010). Tempered Markov Chain Monte Carlo for training of restricted Boltzmann machines. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Sardinia, Italy., pp. 145–152. Cited on pages 38, 116.

Driver, J., Davis, G., Russell, C., Turatto, M., & Freeman, E. (2001). Segmentation, attention and phenomenal visual objects. *Cognition 80*(1-2), 61–95. Cited on pages 145, 147, 148.

Duncan, J., Humphreys, G., & Ward, R. (1997). Competitive brain activity in visual attention. *Current Opinion in Neurobiology 7*(2), 255–261. Cited on pages 143, 148, 150, 152, 168.

Dura-Bernal, S., Wennekers, T., & Denham, S. (2011). Modelling object perception in cortex: Hierarchical Bayesian networks and belief propagation. In *Information Sciences and Systems (CISS), 2011 45th Annual Conference on*, pp. 1 –6. Cited on pages 17, 168, 186.

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience 3 Suppl*, 1184–1191. PMID: 11127836. Cited on page 22.

Eichhorn, J., Sinz, F., & Bethge, M. (2009). Natural image coding in V1: how much use is orientation selectivity? *PLoS Comput Biol 5*(4), e1000336. Cited on page 178.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature 415*(6870), 429–433. Cited on page 12.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences 8*(4), 162–169. Cited on page 12.

Eslami, S. M. A., Heess, N., & Winn, J. (2012). The shape Boltzmann machine: a strong model of object shape. In *IEEE Conference on Computer Vision and Pattern Recognition 2012*. Cited on pages 70, 91, 99, 186.

Farkhooi, F., Strube-Bloss, M. F., & Nawrot, M. P. (2009). Serial correlation in neural spike trains: Experimental evidence, stochastic modeling, and single neuron variability. *Physical Review E 79*(2), 021905. Cited on page 120.

Fazl, A., Grossberg, S., & Mingolla, E. (2009). View-invariant object category learning, recognition, and search: how spatial and object attention are coordinated using surface-based attentional shrouds. *Cognitive Psychology 58*(1), 1–48. PMID: 18653176. Cited on page 166.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex 1*(1), 1–a–47. Cited on pages 1, 46.

ffytche, D. H. (2005). Two visual hallucinatory syndromes. *Behavioral and brain sciences 28*(6), 763–764. Cited on page 61.

ffytche, D. H. (2007). Visual hallucinatory syndromes: past, present, and future. *Dialogues in Clinical Neuroscience 9*(2), 173–189. Cited on page 63.

ffytche, D. H., & Howard, R. J. (1999). The perceptual consequences of visual loss: 'positive' pathologies of vision. *Brain 122*(7), 1247–1260. Cited on pages 63, 86.

ffytche, D. H., Howard, R. J., Brammer, M. J., David, A., Woodruff, P., & Williams, S. (1998). The anatomy of conscious vision: an fMRI study of visual hallucinations. *Nature Neuroscience 1*(8), 738–742. PMID: 10196592. Cited on pages 95, 98.

Finkel, L. H. (2000). Neuroengineering models of brain disease. *Annual Review of Biomedical Engineering 2*(1), 577–606. Cited on page 97.

Fiser, J., Berkes, B., Orban, G., & Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences 14*, 119–130. Cited on pages 4, 16, 106, 109.

Frank, M. J., & Badre, D. (2011). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex*. Cited on page 190.

Frigg, R., & Hartmann, S. (2008). Models in science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2008 ed.). Cited on pages 47, 48.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Comput Biol 4*(11), e1000211. Cited on pages 15, 183.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat Rev Neurosci 11*(2), 127–138. Cited on pages 43, 182, 183.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences 364*(1521), 1211 –1221. Cited on pages 52, 182, 183.

Friston, K. J. (2005). Hallucinations and perceptual inference. *Behavioral and Brain Sciences 28*(06), 764–766. Cited on pages 65, 97.

Friston, K. J. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences 13*(7), 293–301. Cited on pages 17, 43, 134, 136, 182, 183.

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese 159*(3), 417–458. Cited on pages 31, 43, 194.

Földiák, P. (1990). Forming sparse representations by local anti-hebbian learning. *Biological Cybernetics 64*(2), 165–170. Cited on page 124.

Geisler, W. (2003). Ideal observer analysis. *The visual neurosciences*, 825–837. Cited on pages 11, 12, 14.

George, D., & Hawkins, J. (2009). Towards a mathematical theory of cortical microcircuits. *PLoS Comput Biol 5*(10), e1000532. Cited on pages 17, 168, 184, 186.

Gershman, S., Vul, E., & Tenenbaum, J. (2009a). Perceptual multistability as Markov chain Monte Carlo inference. *Advances in Neural Information Processing Systems 22 22*, 611–619. Cited on pages 106, 122.

Gershman, S., Vul, E., & Tenenbaum, J. (2009b). Perceptual multistability as Markov chain Monte Carlo inference. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22*, pp. 611–619. Cited on pages 106, 109, 110, 130, 131, 132, 136, 142, 194.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation 24*(1), 1–24. Cited on pages 106, 131, 132, 136, 139.

Griffiths, T., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin 138*(3), 415–422. Cited on pages 7, 12, 14, 195, 196.

Griffiths, T., Kemp, C., & Tenenbaum, J. (2008). Bayesian models of cognition. *Cambridge handbook of computational cognitive modeling*, 59–100. Cited on page 6.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences 14*(8), 357–364. Cited on pages 51, 193, 196.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: II. feedback, expectation, olfaction, illusions. *Biological Cybernetics 23*(4), 187–202. PMID: 963125. Cited on pages 52, 65.

Grossberg, S. (2000). How hallucinations may arise from brain mechanisms of learning, attention, and volition. *Journal of the International Neuropsychological Society 6*(05), 583–592. Cited on pages 65, 97.

Grossberg, S., & Swaminathan, G. (2004). A laminar cortical model for 3D perception of slanted and curved surfaces and of 2D images: development, attention, and bistability. *Vision Research 44*(11), 1147–1187. Cited on pages 106, 130.

Hawkins, J., Ahmad, S., & Dubinsky, D. (2011). Hierarchical temporal memory including HTM cortical learning algorithms. Technical Report version 0.2.1, Numenta, Inc. Cited on page 185.

Hinton, G., Sallans, B., & Ghahramani, Z. (1998). A hierarchical community of experts. In *Learning in Graphical Models*, Volume 89, pp. 479–494. Kluwer Academic Publishers. Cited on page 33.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation 14*(8), 1771–1800. Cited on page 36.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences 11*(10), 428–434. Cited on page 2.

Hinton, G. E. (2010a). Learning to represent visual input. *Philosophical Transactions of the Royal Society B: Biological Sciences 365*(1537), 177–184. Cited on pages 46, 190.

Hinton, G. E. (2010b). A practical guide to training restricted Boltzmann machines. Technical report UTML TR 2010-003, Department of Computer Science, Machine Learning Group, University of Toronto. Cited on pages 37, 58.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science 268*(5214), 1158–1161. Cited on page 105.

Hinton, G. E., Krizhevsky, A., & Wang, S. D. (2011). Transforming auto-encoders. In T. Honkela, W. Duch, M. Girolami, & S. Kaski (Eds.), *Artificial Neural Networks and Machine Learning - ICANN 2011*, Volume 6791, pp. 44–51. Berlin, Heidelberg: Springer Berlin Heidelberg. Cited on pages 100, 195.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation 18*(7), 1527–1554. Cited on page 55.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science 313*(5786), 504–507. Cited on pages iv, 20, 33, 39, 45, 58, 156.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation 9*(8), 1735–1780. Cited on pages 176, 195.

Hoffman, R. E., & Dobscha, S. K. (1989). Cortical pruning and the development of schizophrenia: a computer model. *Schizophrenia bulletin 15*(3), 477–490. PMID: 2814376. Cited on page 97.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition 108*(3), 687–701. Cited on pages 106, 122, 134, 136, 180, 183.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences 79*(8), 2554 –2558. Cited on pages 19, 22.

Hoyer, P. O., & Hyvärinen, A. (2003). Interpreting neural response variability as Monte Carlo sampling of the posterior. In S. T. S. Becker & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15*, pp. 277–284. Cambridge, MA: MIT Press. Cited on pages 4, 16, 109, 136, 194.

Hupé, J.-M., James, A. C., Girard, P., Lomber, S. G., Payne, B. R., & Bullier, J. (2001). Feedback connections act on the early part of the responses in monkey visual cortex. *Journal of Neurophysiology 85*(1), 134–145. Cited on page 171.

Hyvärinen, A., Hurri, J., & Hoyer, P. O. (2009). *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer. Cited on pages 51, 178, 179, 182.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research 40*(10-12), 1489–1506. Cited on page 145.

Jaramillo, S., & Pearlmutter, B. A. (2007). Optimal coding predicts attentional modulation of activity in neural systems. *Neural Computation 19*(5), 1295–1312. Cited on page 175.

Jarsky, T., Roxin, A., Kath, W. L., & Spruston, N. (2005). Conditional dendritic spike propagation following distal synaptic activation of hippocampal CA1 pyramidal neurons. *Nat Neurosci 8*(12), 1667–1676. Cited on page 188.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). *SciPy: Open source scientific tools for Python*. Cited on page iv.

Jones, M., & Love, B. C. (2011a). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences 34*(04), 169–188. Cited on pages 7, 10, 14.

Jones, M., & Love, B. C. (2011b). Pinning down the theoretical commitments of Bayesian cognitive models. *Behavioral and Brain Sciences 34*(04), 215–231. Cited on pages 7, 10, 195.

Jordan, M. I. (2004). Graphical models. *Statistical Science 19*(1), 140–155. Cited on page 9.

Jordan, M. I. (2009). Bayesian or frequentist, which are you? *Machine Learning Summer School (MLSS), Cambridge 2009; http://videolectures.net/mlss09uk_jordan_bfway/*. Cited on page 9.

Kaas, J. H. (2011). Neocortex in early mammals and its subsequent variations. *Annals of the New York Academy of Sciences 1225*(1), 28–36. Cited on pages 1, 193.

Kang, M.-S., & Blake, R. (2010). What causes alternations in dominance during binocular rivalry? *Attention, Perception, & Psychophysics 72*(1), 179–186. Cited on pages 108, 129, 138, 139, 142, 192.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology 55*(1), 271–304. Cited on pages 9, 10, 11, 50.

Kim, Y.-J., Grabowecky, M., & Suzuki, S. (2006). Stochastic resonance in binocular rivalry. *Vision Research 46*(3), 392–406. Cited on page 141.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science 220*(4598), 671 –680. Cited on page 22.

Kirsch, J. A., Güntürkün, O., & Rose, J. (2008). Insight without cortex: Lessons from the avian brain. *Consciousness and Cognition 17*(2), 475–483. Cited on page 183.

Knapen, T., Kanai, R., Brascamp, J., van Boxtel, J., & van Ee, R. (2007). Distance in feature space determines exclusivity in visual rivalry. *Vision Research 47*(26), 3269–3275. PMID: 17950397. Cited on pages 124, 125.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences 27*(12), 712–719. PMID: 15541511. Cited on pages 6, 12, 17, 136.

Kosslyn, S. M., & Thompson, W. L. (2003). When is early visual cortex activated during visual mental imagery? *Psychological Bulletin 129*(5), 723–746. PMID: 12956541. Cited on page 190.

Lakhtakia, A. (1996). *Models and Modelers of Hydrogen: Thales, Thomson, Rutherford, Bohr, Sommerfeld, Goudsmit, Heisenberg, Schrödinger, Dirac, Sallhofer*. World Scientific. Cited on page 195.

Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences 10*(11), 494–501. Cited on pages 102, 103.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences 23*(11), 571–579. PMID: 11074267. Cited on pages 144, 149, 152, 156, 171.

Larochelle, H., & Bengio, Y. (2008). Classification using discriminative restricted Boltzmann machines. Helsinki, Finland, pp. 536–543. ACM. Cited on page 55.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., & Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. Corvalis, Oregon, pp. 473–480. ACM. Cited on page 174.

Larochelle, H., & Hinton, G. (2010). Learning to combine foveal glimpses with a third-order Boltzmann machine. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23*, pp. 1243–1251. Cited on pages 168, 190.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. In *The handbook of brain theory and neural networks*, Volume 3361, pp. 255–258. Cambridge MA: MIT Press. Cited on pages 170, 188.

Lee, H., Ekanadham, C., & Ng, A. Y. (2008). Sparse deep belief net model for visual area V2. *Advances in Neural Information Processing Systems 20*. Cited on pages 20, 46, 47, 51, 69, 156, 179, 181.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, Quebec, Canada, pp. 1–8. ACM. Cited on pages 100, 170, 188.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A 20*(7), 1434–1448. Cited on pages 2, 3, 4, 7, 16, 17, 18, 65, 91, 105, 137.

Leopold, & Logothetis (1999). Multistable phenomena: changing views in perception. *Trends in Cognitive Sciences 3*(7), 254–264. PMID: 10377540. Cited on pages 105, 125, 128.

Levy, R., Reali, F., & Griffiths, T. (2009). Modeling the effects of memory on human online sentence processing with particle filters. *Advances in neural information processing systems 21*, 937–944. Cited on pages 16, 109.

Likova, L., & Tyler, C. (2008). Occipital network for figure/ground organization. *Experimental Brain Research 189*(3), 257–267. Cited on page 171.

Liu, L., Tyler, C. W., & Schor, C. M. (1992). Failure of rivalry at low contrast: Evidence of a suprathreshold binocular summation process. *Vision Research 32*(8), 1471–1479. Cited on page 108.

Maass, W., & Zador, A. M. (1999). Dynamic stochastic synapses as computational units. *Neural Computation 11*(4), 903–917. Cited on page 113.

MacKay, D. J. C. (2002). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. Cited on pages 22, 42, 46.

Manford, M., & Andermann, F. (1998). Complex visual hallucinations. Clinical and neurobiological insights. *Brain 121*(10), 1819–1840. Cited on pages 64, 66, 90, 95.

Marder, E., & Goaillard, J.-M. (2006). Variability, compensation and homeostasis in neuron and network function. *Nat Rev Neurosci 7*(7), 563–574. Cited on page 67.

Marr, D., Ullman, S., & Poggio, T. (2010). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Mit Press. Cited on page 11.

Mason, O. J., & Brady, F. (2009). The psychotomimetic effects of short-term sensory deprivation. *The Journal of Nervous and Mental Disease 197*(10), 783–785. Cited on page 97.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences 14*(8), 348–356. Cited on page 47.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review 102*, 419–457. Cited on pages 51, 191.

Memisevic, R., & Conrad, C. (2011). Stereopsis via deep learning. In *NIPS workshop on Deep Learning and Unsupervised Feature Learning*. Cited on pages 141, 190.

Meng, M., & Tong, F. (2004). Can attention selectively bias bistable perception? Differences between binocular rivalry and ambiguous figures. *Journal of Vision 4*(7). Cited on pages 121, 122.

Menon, G. J., Rahman, I., Menon, S. J., & Dutton, G. N. (2003). Complex visual hallucinations in the visually impaired: the Charles Bonnet Syndrome. *Survey of Ophthalmology 48*(1), 58–72. PMID: 12559327. Cited on pages 63, 66, 76, 80, 84, 97, 98, 99.

Merabet, L. B., Kobayashi, M., Barton, J., & Pascual-Leone, A. (2003). Suppression of complex visual hallucinatory experiences by occipital transcranial magnetic stimulation: A case report. *Neurocase: The Neural Basis of Cognition 9*(5), 436. Cited on pages 87, 88.

Mesulam, M. M. (1998). From sensation to cognition. *Brain 121*(6), 1013 –1052. Cited on pages 1, 193.

Mohamed, A., Dahl, G. E., & Hinton, G. (2012). Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing 20*(1), 14–22. Cited on page 20.

Montavon, G., Braun, M., Müller, K., & Berlin, T. (2012). Deep Boltzmann machines as feed-forward hierarchies. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. Cited on page 57.

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*. Cited on pages 106, 109, 130, 132, 138.

Morrison, J., & David, A. S. (2005). Now you see it, now you don't : More data at the cognitive level needed before the PAD model can be accepted. *Behavioral and brain sciences (Print) 28*(6), 770–771. Cited on page 61.

Mueser, K. T., Bellack, A. S., & Brady, E. U. (1990). Hallucinations in schizophrenia. *Acta Psychiatrica Scandinavica 82*(1), 26–29. PMID: 2399817. Cited on page 61.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics 66*(3), 241–251. Cited on page 52.

Mumford, D. (1994). Neuronal architectures for pattern-theoretic problems. In C. Koch & J. Davis (Eds.), *Large-scale neuronal theories of the brain*, pp. 125–152. Cambridge MA: MIT Press. Cited on page 105.

Murray, J. F., & Kreutz-Delgado, K. (2007). Visual recognition and inference using dynamic overcomplete sparse learning. *Neural Computation 19*(9), 2301–2352. Cited on pages 17, 168, 185, 186.

Murray, M. M., Foxe, D. M., Javitt, D. C., & Foxe, J. J. (2004). Setting boundaries: Brain dynamics of modal and amodal illusory shape completion in humans. *J. Neurosci. 24*(31), 6898–6903. Cited on page 192.

Murray, S. O., Schrater, P., & Kersten, D. (2004). Perceptual grouping and the interactions between visual cortical areas. *Neural Networks 17*(5–6), 695–705. Cited on page 180.

Nair, V., & Hinton, G. (2009). 3D object recognition with deep belief nets. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22*, pp. 1339–1347. Cited on pages 20, 51, 69, 179.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In J. Fürnkranz & T. Joachims (Eds.), *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, Haifa, Israel, pp. 807–814. Omnipress. Cited on page 132.

Nishimoto, S., Vu, A., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology 21*(19), 1641–1646. Cited on page 56.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature 381*(6583), 607–609. Cited on pages 16, 47, 178.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research 37*(23), 3311–3325. Cited on pages 99, 178.

Olson, C. R. (2001). Object-based vision and attention in primates. *Current Opinion in Neurobiology 11*(2), 171–179. Cited on page 146.

Osindero, S., & Hinton, G. (2008). Modeling image patches with a directed hierarchy of Markov random fields. *Advances in Neural Information Processing Systems 20*, 1121–1128. Cited on pages 100, 187, 190.

Ostrovsky, Y., Meyers, E., Ganesh, S., Mathur, U., & Sinha, P. (2009). Visual parsing after recovery from blindness. *Psychological Science 20*(12), 1484–1491. Cited on page 171.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., & Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology 10*(1), e1001251. Cited on page 56.

Perry, E. K., & Perry, R. H. (1995). Acetylcholine and hallucinations - disease-related compared to drug-induced alterations in human consciousness. *Brain and Cognition 28*(3), 240–258. Cited on pages 66, 90.

Peterson, M. A., & Gibson, B. S. (1991). Directing spatial attention within an object: Altering the functional equivalence of shape descriptions. *Journal of Experimental Psychology: Human Perception and Performance 17*(1), 170–182. Cited on pages 121, 122.

Plaut, D., & McClelland, J. (2010a). Postscript: Parallel distributed processing in localist models without thresholds. *Psychological Review 117*(1), 289–290. Cited on page 51.

Plaut, D. C., & McClelland, J. L. (2010b). Locating object knowledge in the brain: Comment on Bowers's (2009) attempt to revive the grandmother cell hypothesis. *Psychological review 117*(1), 284–290. Cited on page 51.

Plummer, C., Kleinitz, A., Vroomen, P., & Watts, R. (2007). Of roman chariots and goats in overcoats: The syndrome of Charles Bonnet. *Journal of Clinical Neuroscience 14*(8), 709–714. Cited on pages 63, 66, 67, 95, 99.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology 32*(1), 3–25. Cited on pages 150, 151.

Qiu, F. T., Sugihara, T., & von der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience 10*(11), 1492–1499. Cited on page 149.

Ranzato, M., Krizhevsky, A., & Hinton, G. E. (2010). Factored 3-way restricted Boltzmann machines for modeling natural images. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Volume 9, pp. 621–628. Cited on pages 46, 100, 188.

Ranzato, M., Susskind, J., Mnih, V., & Hinton, G. (2011). On deep generative models with applications to recognition. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2857–2864. IEEE. Cited on page 100.

Rao, R., & Ballard, D. (2004). Probabilistic models of attention based on iconic representations and predictive coding. In *Neurobiology of Attention.*, pp. 553–561. Academic Press, New York. Cited on page 180.

Rao, R. P. (1999). An optimal estimation approach to visual perception and learning. *Vision Research 39*(11), 1963–1989. Cited on pages 14, 17, 168, 180.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience 2*(1), 79–87. PMID: 10195184. Cited on pages 2, 16, 52, 65, 99, 168, 179.

Rao, R. P. N., & Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation 9*(4), 721–763. Cited on pages 99, 179, 180, 194.

Reichert, D., Seriès, P., & Storkey, A. (2010). Hallucinations in Charles Bonnet syndrome induced by homeostasis: a deep Boltzmann machine model. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23*, pp. 2020–2028. Cited on pages 5, 62.

Reichert, D. P., Seriès, P., & Storkey, A. J. (2011a). A hierarchical generative model of recurrent object-based attention in the visual cortex. In T. Honkela, W. Duch, M. Girolami, & S. Kaski (Eds.), *Artificial Neural Networks and Machine Learning - ICANN 2011*, Volume 6791, pp. 18–25. Berlin, Heidelberg: Springer Berlin Heidelberg. Cited on pages 5, 144.

Reichert, D. P., Seriès, P., & Storkey, A. J. (2011b). Neuronal adaptation for sampling-based probabilistic inference in perceptual bistability. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24*, pp. 2357–2365. Cited on pages 5, 107.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition 7*(1), 17. Cited on pages 143, 144, 146, 148, 150, 152, 153, 173.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron 61*(2), 168–185. PMID: 19186161. Cited on page 147.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience 2*(11), 1019–1025. PMID: 10526343. Cited on pages 170, 186, 188.

Rifai, S., Bengio, Y., Dauphin, Y., & Vincent, P. (2012). A generative process for sampling contractive auto-encoders. In J. Langford & J. Pineau (Eds.), *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, New York, NY, USA, pp. 1855–1862. Omnipress. Cited on page 100.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature 323*(6088), 533–536. Cited on page 20.

Ruppin, E., Reggia, J. A., & Horn, D. (1996). Pathogenesis of schizophrenic delusions and hallucinations: A neural model. *Schizophrenia Bulletin 22*(1), 105–121. Cited on pages 68, 97, 99.

Salakhutdinov, R., & Hinton, G. (2007). Learning a nonlinear embedding by preserving class neighbourhood structure. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Volume 11. Cited on page 73.

Salakhutdinov, R., & Hinton, G. (2009). Deep Boltzmann machines. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Volume 5, pp. 448–455. Cited on pages 20, 39, 40, 42, 45, 56, 100, 156.

Salakhutdinov, R., & Larochelle, H. (2010). Efficient learning of deep Boltzmann machines. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Volume 9, pp. 693–700. Cited on page 156.

Salakhutdinov, R., Tenenbaum, J., & Torralba, A. (2011a). One-shot learning with a hierarchical nonparametric Bayesian model. *CSAIL Technical Reports*. Cited on page 47.

Salakhutdinov, R. R., Tenenbaum, J., & Torralba, A. (2011b). Learning to learn with compound HD models. In J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. N. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24*, pp. 2061–2069. Cited on pages 51, 191.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review 117*(4), 1144–1167. Cited on pages 4, 14, 16, 109, 131, 136, 142, 193, 194, 196.

Santhouse, A. M., Howard, R. J., & ffytche, D. H. (2000). Visual hallucinatory syndromes and the anatomy of the visual brain. *Brain 123*(10), 2055–2064. Cited on pages 63, 98.

Sarter, M., Hasselmo, M. E., Bruno, J. P., & Givens, B. (2005). Unraveling the attentional functions of cortical cholinergic inputs: interactions between signal-driven and cognitive modulation of signal detection. *Brain Research. Brain Research Reviews 48*(1), 98–111. PMID: 15708630. Cited on pages 90, 96.

Saxe, A. M., Bhand, M., Mudur, R., Suresh, B., & Ng, A. Y. (2011). Unsupervised learning models of primary cortical receptive fields and receptive field plasticity. In J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. N. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24*, pp. 1971–1979. Cited on pages 47, 181.

Scholl, B. J. (2001). Objects and attention: the state of the art. *Cognition 80*(1-2), 1–46. Cited on pages 146, 148, 149, 166.

Schultz, G., & Melzack, R. (1991). The Charles Bonnet Syndrome: 'phantom visual images'. *Perception 20*(6), 809–825. PMID: 1816537. Cited on pages 63, 66.

Schwitzgebel, E. (2011). Belief. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2011 ed.). Cited on page 11.

Sejnowski, T. (1986). Higher-order Boltzmann machines. In *American Institute of Physics Conference Series*, Volume 151, pp. 398–403. Cited on page 188.

Serences, J. T., & Yantis, S. (2006). Selective visual attention and perceptual coherence. *Trends in Cognitive Sciences 10*(1), 38–45. PMID: 16318922. Cited on pages 144, 146, 150, 152, 153, 167.

Shipp, S. (2004). The brain circuitry of attention. *Trends in Cognitive Sciences 8*(5), 223–230. PMID: 15120681. Cited on pages 171, 191.

Shpiro, A., Moreno-Bote, R., Rubin, N., & Rinzel, J. (2009). Balance between noise and adaptation in competition models of perceptual bistability. *Journal of Computational Neuroscience 27*(1), 37–54. Cited on pages 129, 138, 139, 141, 192.

Sillito, A. M., Cudeiro, J., & Jones, H. E. (2006). Always returning: feedback and sensory processing in visual cortex and thalamus. *Trends in Neurosciences 29*(6), 307–316. Cited on page 171.

Smolensky, P. (1986). Information processing in dynamical systems: foundations of harmony theory. In *Parallel distributed processing: explorations in the microstructure of cognition. Vol. 1. Foundations*, pp. 194–281. Cambridge, MA: MIT Press. Cited on page 27.

Spencer, K. M., & McCarley, R. W. (2005). Visual hallucinations, attention, and neural circuitry : Perspectives from schizophrenia research. *Behavioral and brain sciences (Print) 28*(6), 774. Cited on page 61.

Spratling, M. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research 48*(12), 1391–1408. Cited on page 184.

Spratling, M. W. (2011). Unsupervised learning of generative and discriminative weights encoding elementary image components in a predictive coding model of cortical function. *Neural Computation 24*(1), 60–103. Cited on pages 124, 184.

Spratling, M. W., & Johnson, M. H. (2004). A feedback model of visual attention. *Journal of Cognitive Neuroscience 16*(2), 219–237. Cited on page 149.

Sterzer, P., Kleinschmidt, A., & Rees, G. (2009). The neural bases of multistable perception. *Trends in Cognitive Sciences 13*(7), 310–318. Cited on page 105.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nat Neurosci 9*(4), 578–585. Cited on pages 13, 108.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences 13*(9), 403–409. Cited on page 150.

Sundareswara, R., & Schrater, P. R. (2008). Perceptual multistability predicted by search model for Bayesian decisions. *Journal of Vision 8*(5), 1–19. Cited on pages 15, 106, 109, 110, 121, 130, 136, 137, 142, 189, 192.

Sur, M., & Leamey, C. A. (2001). Development and plasticity of cortical areas and networks. *Nat Rev Neurosci 2*(4), 251–262. Cited on page 1.

Tang, Y., & Eliasmith, C. (2010). Deep networks for robust visual recognition. In *Proceedings of the 27th Annual International Conference on Machine Learning*, Haifa, Israel, pp. 1055–1062. Cited on page 161.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science 331*(6022), 1279 –1285. Cited on page 17.

Teunisse, R. J., Zitman, F. G., Cruysberg, J. R. M., Hoefnagels, W. H. L., & Verbeek, A. L. M. (1996). Visual hallucinations in psychologically normal people: Charles Bonnet's Syndrome. *The Lancet 347*(9004), 794–797. Cited on pages 63, 99.

Thomasson, A. (2012). Categories. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2012 ed.). Cited on page 14.

Tieleman, T. (2008). Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th Annual International Conference on Machine Learning*, Helsinki, Finland, pp. 1064–1071. Cited on pages 38, 111.

Tieleman, T. (2010). Gnumpy: an easy way to use GPU boards in python. Technical Report UTML TR 2010-002, University of Toronto, Department of Computer Science. Cited on page iv.

Tieleman, T., & Hinton, G. (2009). Using fast weights to improve persistent contrastive divergence. In *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, Quebec, Canada, pp. 1033–1040. ACM. Cited on pages 38, 111, 112.

Tiesinga, P., Fellous, J.-M., & Sejnowski, T. J. (2008). Regulation of spike timing in visual cortical circuits. *Nature reviews. Neuroscience 9*(2), 97–107. PMID: 18200026 PMCID: 2868969. Cited on page 188.

Titsias, M., & Williams, C. (2006). Sequential learning of layered models from video. In *Toward Category-Level Object Recognition*, pp. 577–595. Cited on pages 172, 173.

Tong, F., Meng, M., & Blake, R. (2006). Neural bases of binocular rivalry. *Trends in Cognitive Sciences 10*(11), 502–511. Cited on pages 105, 124, 125, 128, 133, 137.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol Bull 215*(3), 216–242. Cited on page 102.

Toppino, T. C. (2003). Reversible-figure perception: mechanisms of intentional control. *Perception & Psychophysics 65*(8), 1285–1295. PMID: 14710962. Cited on pages 121, 122.

Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology 6*(2), 171–178. Cited on page 169.

Tsotsos, J. (2011a). Attention, recognition, and binding. In *A computational perspective on visual attention*. The MIT Press. Cited on page 149.

Tsotsos, J. (2011b). *A computational perspective on visual attention*. The MIT Press. Cited on pages 145, 149, 152, 156, 185.

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences 13*(03), 423–445. Cited on page 149.

Tsotsos, J. K., Rodriguez-Sanchez, A. J., Rothenstein, A. L., & Simine, E. (2008). The different stages of visual recognition need different attentional binding strategies. *Brain Research 1225*, 119–132. Cited on pages 144, 149, 152, 169.

Turrigiano, G. G. (2008). The self-tuning neuron: synaptic scaling of excitatory synapses. *Cell 135*(3), 422–435. PMID: 18984155. Cited on pages 67, 68, 97.

Turrigiano, G. G., & Nelson, S. B. (2000). Hebb and homeostasis in neuronal plasticity. *Current Opinion in Neurobiology 10*(3), 358–364. Cited on pages 67, 76.

van Ee, R., Adams, W. J., & Mamassian, P. (2003). Bayesian modeling of cue interaction: bistability in stereoscopic slant perception. *Journal of the Optical Society of America A 20*(7), 1398–1406. Cited on page 106.

van Ee, R., Noest, A. J., Brascamp, J. W., & van den Berg, A. V. (2006). Attentional control over either of the two competing percepts of ambiguous stimuli revealed by a two-parameter analysis: means do not make the difference. *Vision Research 46*(19), 3129–3141. PMID: 16650452. Cited on page 122.

Vecera, S. (2000). Toward a biased competition account of object-based segregation and attention. *Brain and Mind 1*(3), 353–384. Cited on page 148.

Vecera, S. P., Behrmann, M., Shipley, T. F., & Kellman, P. J. (2001). 6 attention and unit formation: A biased competition account of object-based attention. In *From Fragments to Objects Segmentation and Grouping in Vision*, Volume Volume 130, pp. 145–180. North-Holland. Cited on pages 146, 147.

Vilares, I., & Kording, K. (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences 1224*(1), 22–39. Cited on pages 6, 9, 15, 16, 137.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. Helsinki, Finland, pp. 1096–1103. ACM. Cited on page 20.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? optimal decisions from very few samples. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.* Cited on pages 16, 109.

Wainwright, M. J., & Jordan, M. I. (2007). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning 1*(1–2), 1–305. Cited on page 42.

Ward, L. (2008). Attention. *Scholarpedia 3*(10), 1538. Cited on page 145.

Webb, B. (2009). Animals versus animats: Or why not model the real iguana? *Adaptive Behavior 17*(4), 269–286. Cited on pages 47, 48.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nat Neurosci 5*(6), 598–604. Cited on pages 13, 99.

Welling, M. (2009). Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, Quebec, Canada, pp. 1121–1128. ACM. Cited on page 113.

Welling, M., Rosen-Zvi, M., & Hinton, G. (2005). Exponential family harmoniums with an application to information retrieval. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*, pp. 1481–1488. Cambridge, MA: MIT Press. Cited on page 22.

Whiteley, L. (2008). *Uncertainty, Reward, and Attention in the Bayesian Brain.* Ph. D. thesis, University College London. Cited on pages 15, 150, 151.

Wilson, H. R. (2007). Minimal physiological conditions for binocular rivalry and rivalry memory. *Vision Research 47*(21), 2741–2750. Cited on pages 106, 138.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation 14*(4), 715–770. Cited on page 184.

Younes, L. (1989). Parametric inference for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields 82*(4), 625–645. Cited on page 38.

Yu, A. J., & Dayan, P. (2002). Acetylcholine in cortical inference. *Neural Networks: The Official Journal of the International Neural Network Society 15*(4-6), 719–730. PMID: 12371522. Cited on pages 65, 66, 90, 91, 97, 101, 137, 138, 151.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron 46*(4), 681–692. Cited on pages 137, 151.

Yu, A. J., Dayan, P., & Cohen, J. D. (2009). Dynamics of attentional selection under conflict: toward a rational Bayesian account. *Journal of experimental psychology. Human perception and performance 35*(3), 700–717. PMID: 19485686. Cited on page 151.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences 10*(7), 301–308. PMID: 16784882. Cited on pages 2, 3, 4, 16, 18, 65, 105.

Zhou, G., Sohn, K., & Lee, H. (2012). Online incremental feature learning with denoising autoencoders. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, La Palma, Canary Islands, pp. 1453–1461. Cited on page 191.

Zhou, Y. H., Gao, J. B., White, K. D., Yao, K., & Merk, I. (2004). Perceptual dominance time distributions in multistable visual perception. *Biological Cybernetics 90*(4), 256–263. Cited on pages 119, 120.

Zou, W. Y., Ng, A. Y., & Kai Yu (2011). Unsupervised learning of visual invariance with temporal coherence. In *NIPS workshop on Deep Learning and Unsupervised Feature Learning*. Cited on page 190.